EUROPEAN COMMISSION
DG REFORM

Eesti andmehalduse metoodikaprojekt

Andmekirjelduse juhis

August 2020

Funded by the European Commission's Structural Reform Support Programme and implemented by Ernst & Young Baltic AS in cooperation with the European Commission's Directorate-General for Structural Reform Support

Sisukord

1. Sissejuhatus

Käesolev juhis on üks osa andmehalduse raamistikust,[1] mille eesmärk on aidata organisatsioonil kehtestada ühtseid põhimõtteid, protseduure ning paika seada andmehalduse korraldamiseks vajalikud rollid ja mõõdikud. **Andmehaldus on organisatsiooni võimekus hallata andmeid varana.** Peamine kasu tõhusast andmehalduse korraldustest tekib organisatsioonile läbi andmetest täiendava väärtuse loomise ja kvaliteetsemate juhtimisotsuste langetamise. Andmehalduse raamistik annab organisatsioonile ühtlustatud andmetega töötamise vormi ja struktureeritud ning sujuva suhtlemise viisi erinevate üksuste vahel.

Andmekirjeldus esitab andmestike ja neis sisalduvate andmete kirjelduse (metaandmed). Andmete kirjeldamiseks kasutatavad metaandmed aitavad tagada andmete:

- sisu kirjeldamise ja seletamise;
- ülevaatlikust ja leitavust;
- usaldusväärsust;
- arusaadavust nii tehniliste kui äriliste kasutajate poolt;
- masinmõistetavust ja semantilist koostalitusvõimet;
- elukäigu haldamist.

Süsteemselt hallatud andmekirjeldus annab organisatsioonile ülevaate oma andmetega seotud varadest. Varadeks on andmestikud, andmeteenused, sõnastikud, klassifikaatorid ja loendid ning muud halltavad üksused, kuid peamiselt hallatakse andmeid andmestike kaupa. Andmekirjelduse olemasolu võimaldab andmete leidmist, kasutamist ja neist arusaamist nii inimestele kui ka n-ö masinmõistetavalt infosüsteemidele läbi mõistete ja sõnastike. Andmekirjeldus on andmehalduse närvivõrk, mis aitab asutusel pidada ülevaadet andmevaradest, tagab andmete usaldusväärsuse ja võimaldab neid tõhusalt taaskasutada.

Andmekirjelduse peamine roll andmehalduse raamistikus on tagada asutuse siseselt ja asutuste üleselt teadmine olemasolevatest andmestikest ning andmetest ja nende tähendusest. Andmekirjelduse haldamine on pidev tegevus, mille jaoks kehtestatakse organisatsioonis reeglid ja nõuded ning seda tööd teevad kindlaid rolle täitvad ja tööülesandeid teostavad töötajad. Käesolev juhis kirjeldab terviklikult nii andmekirjelduse loomise kui haldamise tegevusi.

1.1 Juhise koostamine

Juhise koostasid Statistikaameti tellimusel ja Euroopa Komisjoni Struktuursete Reformide Toetusteenistuse (SRSS) poolt rahastatavat projekti „Support for the establishment of data governance services“ (Teotus andmehaldusele. Eesti andmehalduse metoodikaprojekt) teostanud ettevõtte Ernst & Young eksperdid. Juhis koostati 2019 sügisest 2020 aasta suveni ja selle aluseks olid Eestis varem andmehalduse kirjeldamise vallas tehtu ning rahvusvahelised standardid ja parem praktika.

Tellija poolne projektijuht oli Statistikaameti andmehalduse ekspert Veiko Berendsen. Täitja poolne projektijuht oli Siim Aben, eksperdid olid Kuldar Aas ja Raivo Ruusalepp.

[1] vt. EY, *Andmehalduse raamistik* (2020)

1.2 Andmekirjelduse juhise eesmärgid

Juhis on loodud selleks, et asutustel oleks andmekirjelduse loomiseks ja haldamiseks olemas tegevusjuhised, kvaliteedinõuded ja standard. 2019. a. sügisel läbi viidud hetkeolukorra analüüs näitas, et andmete kirjeldamisele on seni asutustes vähe tähelepanud pööratud ja ühtne praktika puudub. Seetõttu on andmekirjelduse juhise peamine eesmärk juhendada organisatsioone andmekirjelduse koostamisel ja haldamisel.

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

1.3 Andmekirjelduse juhise sihtrühmad

Juhise sihtrühmaks on asutused ja organisatsioonid, kes töötlevad andmeid nii andmekogudes kui ka muudes infosüsteemides ja andmebaasides ning kes kasutavad neid andmeid nii esmase infoallikana toiminguteks ja teenusteks, aga ka teiseseks kasutamiseks aruandluses ja statistikas.

Juhis on suunatud eelkõige erialaspetsialistidele – andmehalduritele, kes korraldavad organisatsiooni andmehaldust ning IT spetsialistidele, kes loovad andmeteenuseid ja andmemudeleid.

Andmekirjelduse juhise peamised kasutajad organisatsiooni sees on **andmehaldusega tegelevad spetsialistid**, kes saavad andmekirjelduse kaudu ülevaate organisatsiooni andmestikest kui varadest ning on suutelised korraldama asutuste vahelist andmevahetust; **IT arendajad** ja nende andmespetsialistid, kes saavad ülevaate organisatsiooni andmestikest ning on suutelised tagama tõhusat asutuse infosüsteemide arendamist ja uuendamist; organisatsiooni **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.

Väljaspool organisatsiooni on juhis mõeldud **riigi** keskse **andmehalduse rakenduse halduritele** ja kasutajatele ning andmeportaalide pidajatele, kes saavad ülevaate sellest, milliseid andmeid kus kogutakse ja hoitakse; IT arendajatele, ning laiemale avalikkusele, kes on huvitatud **avaandmetest** ja nende kasutamisest.

1.4 Andmekirjelduse juhise ülesehitus ja põhimõtted

Juhis annab suunised andmekirjelduse protsesside välja töötamiseks ja juurutamiseks organisatsioonis ning vajalike rollide ja vastutuste määramiseks nende protsesside realiseerimisel. Juhis koosneb kolmest põhiosast, mis seletavad mida ja miks on vaja andmekirjeldusega teha (ptk. 2), kuidas andmekirjelduse haldamist teostada (ptk. 3) ning milliseid abivahendeid on seejuures võimalik kasutada (Lisad). Andmekirjelduse protsessid katavad kirjeldatavate andmestike kindlaksmääramist, andmekirjelduse mudeli loomist, sõnastike valimist ning andmete kirjeldamist ja kirjelduse pidevat haldamist. Juhises kasutatakse sõnastike illustreerimiseks Ehitisregistri baasil loodud näiteid (ptk. 3). Juhis on kasutatav koos erinevate andmekirjelduse loomist võimaldavate rakendustega. Juhis võiks edaspidi olla kasutatav elava *on-line* keskkonnana, mis on seotud teadmusbaasiga ja mida saab kestvalt täiendada.

Andmekirjelduse juhise koostamisel lähtuti rahvusvaheliselt tunnustatuimas käsiraamatus „The Data Management Body of Knowledge (DAMA-DMBOK2) (2nd ed., 2017)“ esitatud andmehalduse mudelist.[2] Andmekirjelduse standardi koostamisel tugineti standarditele „Data Catalog Vocabulary (DCAT)“ (ver. 2, 04.02.2020)[3] ja „Data Documentation Initiative (DDI)“ (Lifecycle 3.3, 20.04.2020).[4]

1.5 Andmekirjelduse õiguslik raamistik

Eestis eraldi andmekirjeldust reguleerivat õigusakti ei ole. Samuti ei ole kehtestatud selgeid nõudeid andmekirjelduse haldamisele. Küll aga sisaldub andmekirjelduse aspekte eri tasandi regulatsioonides, sh. andmekogu-põhised õigusaktid, riigi infosüsteemi ja selle koostalitusvõimet reguleerivad õigusaktid ning raamistikud ja soovituslikud juhised eri asutustelt nagu Riigi Infosüsteemide Amet, Statistikaamet või Rahvusarhiiv. Asutustes on konkreetsete andmekogude kirjeldamise nõudeid esitatud andmekogu (tehnilises) dokumentatsioonis. Ülevaade õigusaktidest, soovituslikest juhistest ja standarditest on juhise Lisas 1.

1.6 Juhises kasutatud lühendid ja mõisted

Termin	Määratlus ja selgitus
Andmed	Informatsiooni taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötluseks [ISO/IEC 2382]
Andmeelelement	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	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 [AvTS §41[1]]

[2] https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf
[3] https://www.w3.org/TR/vocab-dcat-2/
[4] https://ddialliance.org/Specification/DDI-Lifecycle/3.3/

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	Teadmisüksus, 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 eelisterminid ja nende sünonüümid [ISO 25964-1]
Ärimõiste	Organisatsiooni tegevust kirjeldav oskussõna
Ärisõnastik	Ärisõnastik on organisatsioonis kasutatava oskussõnavara ja nende sõnaseletuste loend, mis fikseerib organisatsiooni terminoloogia

2. Andmekirjelduse põhimõtted

2.1 Andmekirjelduse ja metaandmete mõiste

Eri valdkondades ja kontekstides tähendavad andmed erinevaid asju. Arvutisse salvestatud andmetest aru saamiseks tuleb neid alati tõlgendada läbi nende konteksti ja eesmärgi, et vältida vääritimõistmist. Andmete rikastamine kirjeldusega annab neile vajaliku konteksti, et saaksime neid mõista tähendust omava informatsioonina. Andmed saavad tähenduse läbi oma metaandmete. 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 oskame nii meie kui tarkvarasüsteemid kasutada andmeobjekti „pind“ õiges tähenduses.

Andmekirjeldus on struktureeritud metaandmete hulk, milles kirjelduselementide ja nendevaheliste seoste kaudu esitatakse kirjeldatava objekti tähendust väljendavat informatsiooni. Andmekirjelduse objektiks on andmestik ja andmeelement ning mõisted, mis seovad andmeelemente tähendust omavateks andmeobjektideks.

Metaandmete kaudu saame vastuse tüüpilistele küsimustele andmete kohta: kes?, mida?, millal?, kus? ja miks? Andmete kohta käivate metaandmete st andmekirjelduse haldus peab olema korraldatud ja järjepidev, et tagada andmete usaldusväärsuse püsimine ja põlvnemise tuvastatavus (*data lineage)*. Metaandmete haldus hõlmab reeglite kehtestamist ja tööprotsesside rakendamist, et kogu organisatsioonil oleks võimalik andmekirjeldust kasutada, jagada, linkida, analüüsida ja töödelda.

Täpsed metaandmed s.o 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 üldistamise või rühmitamise moel. Nendeks on andmete loogilised ja mõistelised mudelid ning andmestikud (andmekogud jt inforessursid). Samuti võivad mõnel juhul, nagu näiteks klassifikaatorites, olla kirjeldatud kõik andmeväljad.

Metaandmete kaudu avaldub andmete väärtus varana,[5] sest andmekirjelduse mõistete ja neid koondavate sõnastike kaudu on organisatsioonil võimalik saada vastused olulisetele küsimustele, nagu:

- Mis andmed meil olemas on ja mille kohta?
- Kust need andmed pärinevad?
- Kus andmed praegu asuvad?
- Kas andmed on ajakohased?
- Kuidas on andmed muutunud nende kogumisest alates?
- 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. „Metaandmete keel“ võimaldab andmete ning tööprotsesside ja teenuste koostoimet sellistes olukordades nagu andmete sisestamine, leidmine, arvutused ja muul viisil kasutamine nii inimeste kui tarkvara jaoks. Andmemudelid, andmeteenused, andmekvaliteet, andmete avaldamine ja andmete elukäigu haldus kasutavad kõik metaandmeid. Kvaliteetne andmekirjeldus parandab oluliselt ülevaate saamist andmetest, andmestikest ja loob parema aluse põhiandmete selliseks määratlemiseks, mis ületab ühe andmekogu piire ja võimaldab neid kui usaldusväärset allikat taaskasutada riigi erinevates andmekogudes, vähendades seekaudu andmete kogumise dubleerimist.

[5] vt. ka *Andmehalduse raamistik*, ptk. 3.3

2.2 Andmekirjelduses kasutatavad sõnastikud

Andmekirjeldus hõlmab metaandmeid andmestike ja neis sisalduvate andmete kohta. Andmekirjelduses on vaja andmete tähistused, mis tihti on lühendid või akronüümid (näiteks relatsioonilises andmebaasis võib veergude tähisteks olla „jt28“, „akpv“, vmt.), seletada lahti sõnadega, mis on üldarusaadavad. Tähendust selgitavad sõnad esitavad mõisteid terminitena. **Mõiste** on inimese peas olev ettekujutus (teadmusüksus) mingist objektist või nähtusest, **termin** aga on sõnavaraüksus – sõna või sõnaühend, millega mõistet tähistatakse (nt. sõnastikus). Terminite valikust sõltub sisuline andmetest arusaadavus. Mõistet tähistav sõna peab omama üldteadaolevalt või kindlas valdkonnas mitmetele kasutajatele ühesugust tähendust. Et andmete kirjeldamiseks valitud sõnad ei oleks juhuslikud, vaid kokkuleppelised ja laialt arusaadavad, on tuleb sõnadena kasutada termineid ehk oskussõnu. Sellistest sõnadest moodustub märksõnastik, mida andmekirjelduses nimetatakse andmesõnastikuks (*data dictionary*). Sõnastik võib olla esitatud:

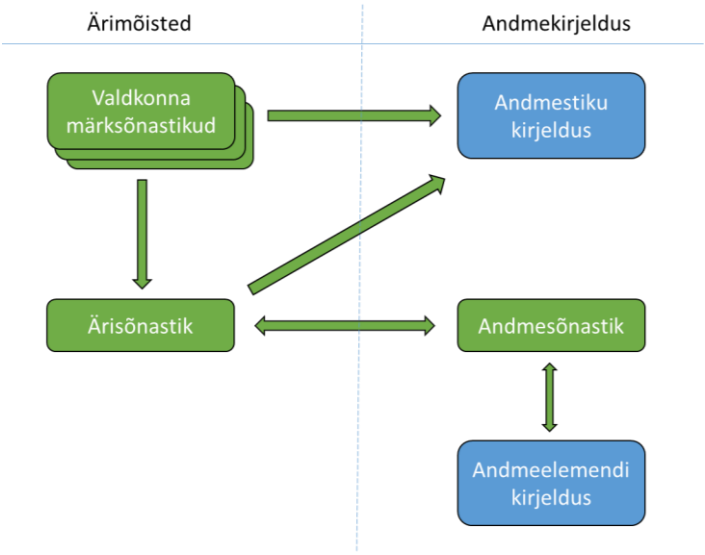
- terminiloendina koos sõnaseletustega, mis määratlevad neid termineid mõne valdkonna jaoks;
- taksonoomiana, kus mõisted on liigitatud hierarhilistesse seostesse ja võivad olla varustatud sõnaseletustega;
- ontoloogiana, mis liigitab valdkonna teadmusüksuseid klassidesse, esitab nendevahelised seosed ja nende omadusi.

Lihtne sõnaseletustega märksõnastik määratleb kasutusel olevad terminid, kuid enamasti ei paku rohkem informatsiooni nendevaheliste seoste ja valdkonna teadmuse „arhitektuuri“ kohta, mida terminid esindavad. Taksonoomiad sisaldavad vähemalt ülem- ja alam-mõiste seoseid, aga võivad olla keerukamatel seostel põhinevad hierarhiad, mis aitavad kiirendada otsinguid ning mõista terminite tähendust ja põlvnemist. Ontoloogiad kirjeldavad ja esitavad mõne valdkonna teadmust masinloetaval kujul, ontoloogia keeles (nt. RDF, OWL). Sõnastike kasutamine andmekirjelduses aitab kaasa kirjelduse ühetaolisusele ja arusaadavusele ning on eelduseks kirjelduse masinmõistetavusele ja andmeteaduse meetodite rakendamisele. Seetõttu ongi selle juhise fookus andmete mõistetavaks tegemine sõnastike abil.

2.2.1 Andmekirjelduse sõnastike tüübid

Andmekirjelduse koostamisel ja haldamisel kasutatakse kolme tüüpi sõnastikke (vt. ka Joonis 1 ja **Error! Reference source not found.**):

- **valdkonna märksõnastik** (*controlled vocabulary*)
- **ärisõnastik** (*business glossary*)
- **andmesõnastik** (*data dictionary*)



Joonis 1. Andmekirjelduse ja ärimõistete sõnastike üldine mudel.

Valdkonna märksõnastik on märksõnade korrastatud loetelu, kus tuuakse välja valdkonna terminite ja nende seoste kirjeldused. Valdkonna märksõnastikud on kasutusel valdkonda hõlmavate teadmiste ja mõistete piiritlemiseks ning nende kokkuleppeliste seletuste esitamiseks. Valdkondlikke 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 valdkondade ülene üldine märksõnastik (EMS).[6] See on kõiki ainevaldkondi hõlmav tesauruse struktuuriga märksõnastik, mida seni on rakendatud peamiselt raamatute, artiklite ja muude teavikute eestikeelseks märksõnastamiseks ja infootsinguks. Kuna selle 61000 terminit on jaotatud 60ks valdkonnaks, siis sobib see kasutamiseks ka kas valdkondliku sõnastikuna või selle loomise alusena. EMS sisaldab ajamärksõnu ja kohanimesid, kuid ei sisalda isikute, asutuste ega organisatsioonide nimesid. EMS toetab masinalt-masinale päringuid MARC-XML vormingus.

Valdkondlikud märksõnastikud võivad olla esitatud ka klassifikaatoritena, valitsemisfunktsioonidena (näiteks VFK, Eesti Majanduse Tegevusalade Klassifikaator (EMTAK/NACE)), õiguslike taksonoomiatena (nt. EuroVoc) või muu taolise valdkondliku jaotusena. Valdkonna märksõnastiku näiteks hariduse valdkonnas Eestis on EstCore2.[7] Lisaks terminitele ja nendevahelistele seostele toob masinloetav versioon välja ka andmeelementide kirjeldused kasutades `owl:DatatypeProperty` kirjeldusi ja nende seosed mõistetega. EstCore2 on seotud ka teiste, üldisemate märksõnastikega (https://schema.org/ jt.).

Ärisõnastik on organisatsioonis kasutatava oskussõnavara ja nende sõnaseletuste loend, mis fikseerib organisatsiooni terminoloogia. Ärisõnastik määratleb organisatsiooni tegevusega seotud mõisted. Ärisõnastik on reeglina tesauruse struktuuriga märksõnastik, 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 ning kehtestada andmetele mõttekaid kvaliteedireegleid. Ärisõnastiku koostab ja seda haldab organisatsiooni äripool ja selle kasutusala on reeglina laiem kui üksnes andmekirjelduse toetamine.

[6] https://ems.elnet.ee/index.php

[7] https://schema.edu.ee/ ja vastav masinloetav sõnastik
https://github.com/hitsa/estcore2/blob/master/ontologies/haridus.rdf

Andmesõnastik on ühest küljest 1) andmemudelis kasutatavate andmeobjektide ja andmeelementide kirjelduste loetelu ja teisest küljest 2) andmekirjelduse käigus tekkiv terminiloend. Andmesõnastik kirjeldab andmeid organisatsiooni tegevuse mõistetena, aga sisaldab lisaks ka andmete kasutamiseks vajalikke metaandmeid (nt. andmetüüp, andmestruktuuride kirjeldus, juurdepääsupiirangud, jmt.). Andmesõnastiku sisu saadakse enamasti andmebaasi füüsilisest mudelist ja seotakse ärisõnastiku terminitega.

2.3 Andmekirjelduse koosseis

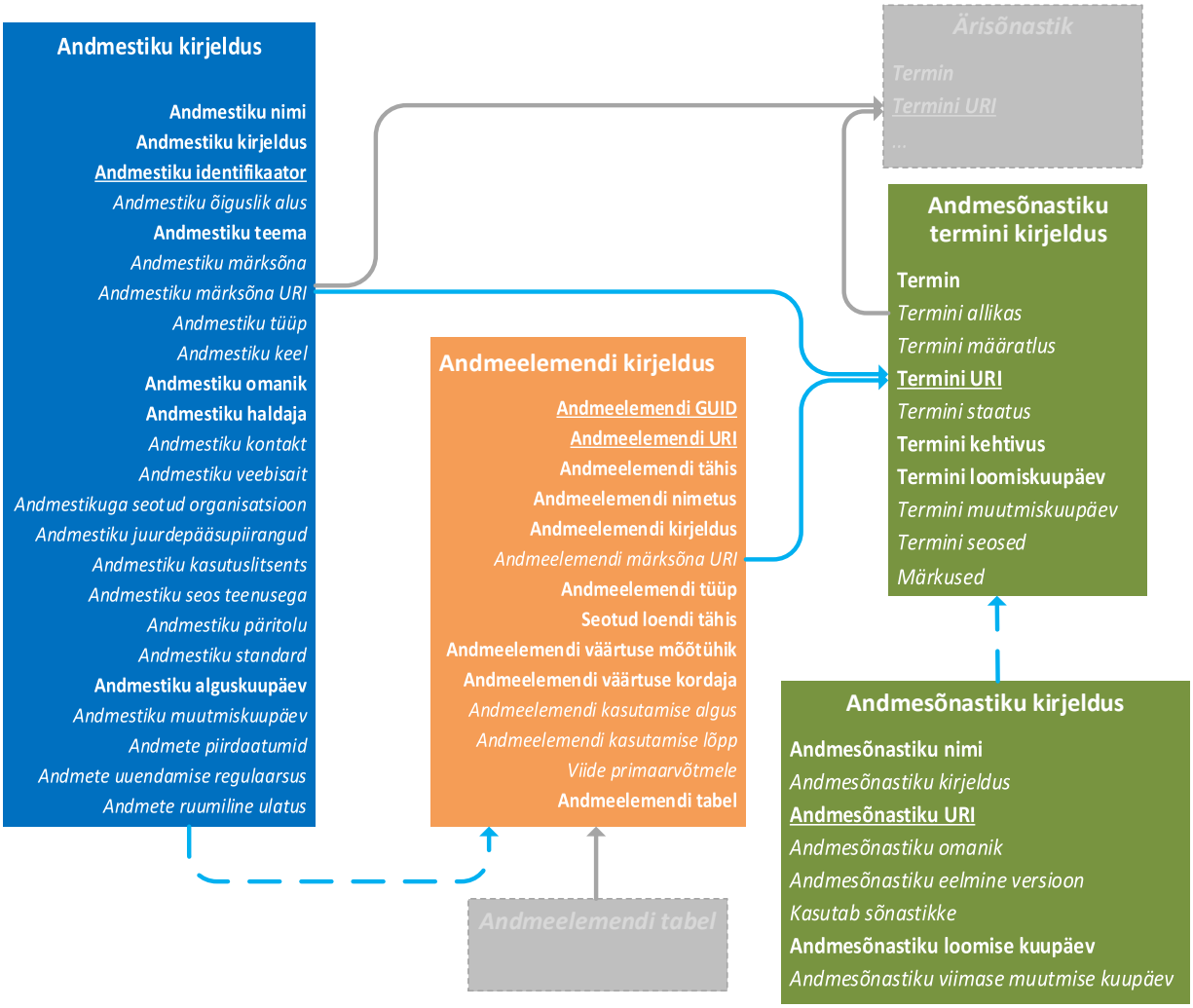
Andmekirjeldus koosneb kahest peamisest komponendist: andmestiku kui terviku kirjeldus ja üksikute andmeelementide kirjeldus. Andmeelementide kirjeldused on andmesõnastiku kaudu seotud organisatsiooni tegevust kirjeldavate mõistetega (ärisõnastikuga).

Andmestiku kirjelduse aluseks on Eestis rahvusvaheline standard „Andmekataloogi sõnastik“ (Data Catalog Vocabulary – DCAT),[8] mille nõutavad kirjelduselemendid on toodud andmekirjelduse standardis (Lisa 2). Andmestiku kirjelduse eesmärgiks on tagada andmestiku kui terviku leitavus ja taaskasutatavus. Andmestiku kirjeldus sisaldab üldist kontekstiinfot selle loomise, omaniku ja haldaja kohta, selle kasutusvõimalusi (s.h. andmestiku kasutamist võimaldav veebisait, kehtivad juurdepääsupiirangud). Vajadusel saab kirjeldada andmestikus sisalduvate andmete geograafilist ulatust (näiteks Tallinna linn, kogu Eesti) või andmete piirdaatumeid (näiteks „andmed alates 1995. aastast“). Andmestiku kirjeldus on suuresti samane juba praegu Riigi infosüsteemi haldussüsteemis (RIHA) andmestike kohta talletatavale infole. Avaldatud andmestiku kirjeldus aitab teistel organisatsioonidel ja kasutajatel otsustada, kas eri andmestike andmed on kattuvad või näiteks ajas ja ruumis erinevad.

Andmeelementide kirjeldamiseks kasutatakse kohandatud „Andmete dokumenteerimise algatuse“ (Data Documentation Initiative – DDI) standardit,[9] mille nõutavad kirjelduselemendid on toodud andmekirjelduse standardis (Lisa 2). Andmeelementide kirjelduse peamiseks eesmärgiks on andmete sisemise struktuuri kirjeldamine selliselt, et need oleks üheselt mõistetavad ja taaskasutatavad nii inimese kui tarkvara poolt. Andmeelemendi, mis üldjuhul on relatsioonilise andmebaasi korral tabeli veerg, kohta talletatakse nii andmemudelist saadud tehnilised andmed kui lisatakse inimloetav nimetus, lühikirjeldus ja kasutusele võtmise aeg. Võimalusel viidatakse ka andmeelemendi väärtuste aluseks olevale klassifikaatorile või loendile.

[8] Data Catalog Vocabulary (Andmekataloogi sõnastik), versioon 2 (2020) https://www.w3.org/TR/vocab-dcat-2/

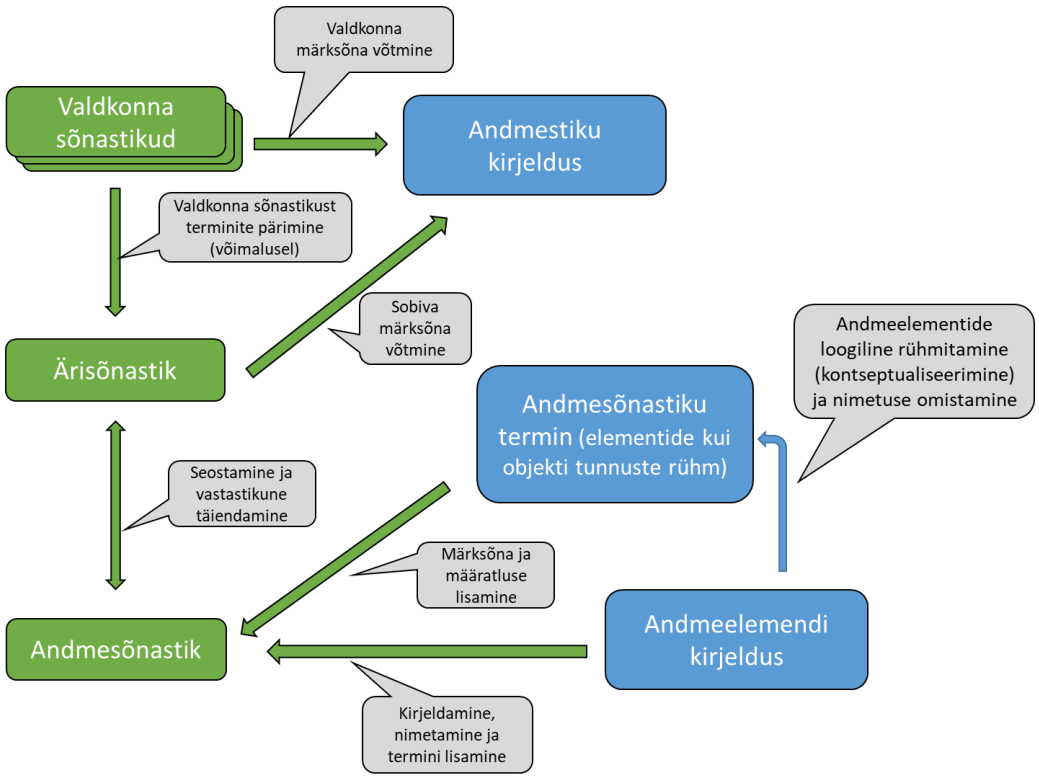
[9] Data Documentation Initiative (Andmete dokumenteerimise algatus), versioon Lifecycle 3.3 (2020) https://ddialliance.org/Specification/DDI-Lifecycle/3.3/



Joonis 2. Andmekirjelduse seosed sõnastikega. Halli värviga esitatud olemid on andmekirjelduse välised, punktiirjoonega on esitatud ilmutamata kujul esinevad alluvusseosed.

2.4 Sõnastike seos andmekirjeldusega

Andmesõnastik on andmestiku füüsilise mudeli (andmeelementide) ja ärisõnastiku ühenduslüli ehk niinimetatud loogiline andmemudel, ainult ilma mudelile omaste seosteta. Andmesõnastik sisaldab sarnaselt ärisõnastikule termineid, mis on üldjuhul võetud ärisõnastikust ja seovad kokku ühte mõistet väljendavad andmeelemendid. Andmestiku ja andmeelemendi kirjelduse seoseid sõnastikega esitab **Error! Reference source not found.**.



Joonis 3. Sõnastike vahelise infovahetuse mudel andmekirjelduses.

Näiteks mõiste „isik“ koosneb andmestikes tavaliselt vähemalt andmeelementidest „eesnimi“, „perekonnanimi“ ja „isikukood“. Sellisena on mõiste „isik“ ka andmeobjekt. Andmesõnastik on oma olemuselt lihtne terminite loend ja selles ei ole terminitel kindlat ärikonteksti. Andmesõnastiku termini määratlus lähtub eelkõige termini sõnalisest tähendusest ja vähem kontekstist. Läbi ühenduse ärisõnastikuga tagatakse, et ühelt poolt on ärisõnastiku terminid seotud kohaste andmeelementidega ning teisalt on olulistele andmeelementidele omistatud kontekst ärisõnastiku termini kaudu. Juhul kui ärisõnastiku terminid on seotud tööprotsesside ja teenustega, on andmesõnastiku kaudu loodud ühenduse kaudu võimalik rakendada andmekvaliteedi reegleid kõikidele andmestikele, kus reegliga kaetud andmeelement esineb.

2.5 Andmekirjelduse haldamine

Andmekirjeldusest kasu saamine eeldab organisatsioonilt teadlikku kirjelduse haldamist, selleks reeglite paika panemist, rollide määramist ja kirjeldusega tegelevatele töötajatele vastavate tööülesannete andmist. Andmekirjelduse juhis (ptk. 3) käsitleb organisatsiooni tasandil terviklikult ülesandeid andmekirjelduse haldamise jaoks – millised tingimused tuleb selleks luua ning andmehalduri tegevusi andmekirjelduse loomisel ja pideval täiendamisel. Andmekirjelduse haldamine ja pidev metaandmete kvaliteedi parendamine tõstavad andmete väärtust. Andmete käsitlemine organisatsiooni varana hõlmab olulise andmehalduse komponendina ka andmekirjeldust, mis omakorda toetab andmepõhist juhtimist ja andmete usaldusväärsuse tuvastamist ning andmete taaskasutamist.

Lihtsamaid andmestikke ja väikest hulka andmeid võib kirjeldada käsitsi tehtud tabelites, kasutades käepäraseid tabelarvutuse vahendeid. Suurema arvu andmeelementide kirjelduste, mahukate sõnastike ja nendevaheliste keerukate seoste käsitsi haldamine ei ole praktiline. Reeglina võetakse andmekirjelduse haldamiseks kasutusele spetsiaalne töövahend. Kõige sagedamini kannab see nimetust andmekataloog, kuid andmekataloogi tarkvararakendusi on mitmeid ja erineva funktsionaalsusega.

Andmekataloog on andmekirjelduse koostamise ja pidamise vahend, mis hoiab andmesõnastiku seoseterviklust andmeelementide ja ärisõnastikuga ning paljudel juhtudel võimaldab hallata ka andmekvaliteedi reegleid. Andmekataloog võib olla ka andmestikele vahetu juurdepääsu vahendiks ja sel juhul on tema funktsionaalsus andmekirjeldusest ulatuslikum.

Andmekirjelduse juhis ei eelda andmekirjelduse haldamist ühegi konkreetse töövahendiga. Juhise koostamisega samaaegselt viisid Riigi Infosüsteemide Amet ja Statistikaamet koostöös Majandus- ja Kommunikatsiooniministeeriumiga läbi analüüsi ja prototüübi arenduse projekti andmekirjelduse haldamise vahendi loomiseks asutustele, projektinimega RIHAKE. Selles kasutatav andmekirjelduse mudel tugineb käesolevas juhises toodud andmekirjelduse standardile (vt. Lisa 2). Kava kohaselt täidaks RIHAKE asutuses andmekirjelduse töövahendi rolli ja selles saab hallata nii andmesõnastikku kui ka andmestiku ja andmeelementide kirjeldusi ning selles on olemas funktsionaalsus asutuses loodud andmekirjelduste edastamiseks RIHAsse ja teistele seotud osapooltele. Kuna juhise valmimise hetkel seda rakendust veel välja arendatud ei ole, on juhise peatükkides toodud näited loodud tavaliste tabelarvutusvahenditega. Juhise koostamisel on arvestatud võimalusega, et asutus haldab oma andmekirjeldust edaspidi RIHAKEses. Samuti arendab Majandus- ja Kommunikatsiooniministeerium uut avaandmete portaali, milles on aluseks käesoleva juhise andmekirjeldusstandard ning DCAT standard.

Ärisõnastike haldamise vahendeid on kirjeldatud *Andmehalduse raamistiku* peatükis 5.2, andmekirjelduse haldamise tarkvarade lühiülevaade on toodud käesoleva juhise Lisas 1.

3. Andmekirjelduse juhis

Andmekirjelduse juhis käsitleb kahte tegevuste tsüklit: 1) andmekirjelduse haldamise üldised tegevused; ning 2) andmekirjelduse loomise ja pideva täiendamise praktilised toimingud, mis on koondatud peatükki 3.5. Esimene neist tagab organisatsiooni valmiduse andmekirjelduse kestvaks parendamiseks ja taaskasutamiseks, teine keskendub kvaliteetse andmekirjelduse loomisele.

Andmekirjelduse haldamine hõlmab käesolevas juhises kogu andmestike ja andmeelementide kirjeldamise ning sõnastikega seotud tegevuste komplekti. See algab andmete kirjeldamise eesmärkide püstitamisest ja lõpeb andmekirjelduse taaskasutusega (nt. edastamine RIHAsse) ning sidumise organisatsiooni teenuste nõuetega. Andmekirjelduse halduse tegevused on kokkuvõtvalt (vt. Joonis 4):

1) Kirjeldatavate andmestike ja teiste varade väljaselgitamine ja määratlemine (objektide piiritlemine)
2) Organisatsiooni jaoks sobivate ja juba kasutusel olevate märksõnastike kaardistamine
3) Organisatsiooni jaoks sobiva andmekirjelduse mudeli ja haldamise arhitektuuri määratlemine
4) Rollide määratlemine andmekirjelduse loomisel ja andmekirjelduse haldamisel
5) Andmekirjelduse loomine, sh sõnastike koostamine, töötamine andmekirjelduse töövahendiga
6) Andmekirjelduse edastamine ja avaldamine
7) Andmekirjelduse sidumine organisatsiooni tööprotsesside ja teenuste kvaliteedi nõuetega.

Joonis 4. Andmekirjelduse haldamise tegevused.

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 rahvusvaheliste vastavate standarditega ja toetab tõhusat andmevahetust.
- Andmekirjeldus on hallatud ning andmekirjeldus ja andmestikud andmekirjelduse töövahendi kaudu üles leitavad 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.

3.1 Andmekirjelduse objektide piiritlemine

Andmekirjelduse koostamise esimeseks eelduseks on asutuse andmetest ülevaate loomine, andmete väärtuse määramine ja kirjeldatavate andmestike (nt. andmekogu, register) valimine. Andmekirjelduse täpsemaks piiritlemiseks kaardistatakse igas valitud andmestikus sisalduvaid komponente (andmebaasid, tabelid, andmeelemendid).

Joonis 5: Andmekirjelduse objektide piiritlemise tegevused

Organisatsioonis leidub üldjuhul mitmeid erineva olulisuse, taaskasutus- ja säilitusväärtusega andmeid. Ühelt poolt hallatakse põhitegevuse jaoks vajalike 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 organisatsiooni kultuuri mõttes kuid mitte laiema andmete taaskasutus- ja säilitusväärtuse vaatest.

Kõigi andmete haldamine, sealhulgas põhjalik kirjeldamine, ei ole organisatsiooni ressursside vaatest alati mõistlik. Samuti on terve andmemassiivi haldamine üldjuhul üle jõu käiv ülesanne. Sellest tulenevalt on andmekirjeldamiseks valmistumisel oluline esimese sammuna saada ülevaade sisuliselt kokku kuuluvatest andmehulkadest ehk andmestikest ning määrata andmed mille kirjeldamine on organisatsiooni vaatest kõige olulisem.

Andmete olulisus sõltub suurel määral iga konkreetse organisatsiooni tegevusvaldkonnast ja ülesehitusest. Seega on olulisuse määramiseks raske välja tuua täpseid ja universaalseid reegleid. Küll on olemas mõningad üldisemad näpunäited:

- **Taaskasutus:** mida rohkem on andmete taaskasutusest huvitatud sisemisi või välimisi kliente, seda mõistlikum on neid andmeid ka põhjaliku andmekirjeldusega varustada.
- **Põhitegevused vs tugitegevused:** üldjuhul on organisatsiooni põhitegevuse käigus tekkinud andmetel oluliselt suurem tähtsus kui tugitegevuste käigus tekkinud andmetel.
- **Säilitusväärtus:** andmete säilitusväärtus (säilitustähtaeg) on üldjuhul määratud lähtuvalt taaskasutusvajaduste ja organisatsiooni tegevuste analüüsi tulemusena, seega on see heaks indikatsiooniks andmete olulisusest.

Olulisuse järgi grupeeritud andmed tuleb omakorda defineerida selgepiiriliste andmestikena. Andmestiku sobiva piiritlemise eesmärgiks on defineerida kirjeldamiseks 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. Tuleb küll aru saada, et alati pole andmestiku piiritlemisel ühte ja ainsat „õiget“ lahendust. Näiteks Ehitisregistrit on võimalik defineerida kui ühte andmestikku (s.o. ehitisregistri andmestik), või kui mitut erinevat andmestikku (ehitise põhiandmete andmestik, järelevalve andmestik, energiamärgise andmestik, jne.).

Kirjeldatava andmestiku valimisel on oluline andmehalduri koostöö sisuvaldkonna spetsialistidega (näiteks andmeomanik, infosüsteemi peakasutaja) (vt. ka *Andmehalduse raamistik*, ptk. 3.3). Andmestiku komponentide (andmebaasid, tabelid, andmeelemendid) analüüsimisel tuleb konsulteerida ka tehniliste halduritega (näiteks andmebaasi administraator).

Andmestiku piiritlemisel tuleks tähele panna:

- 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 kui andmekogu alam-andmestikud. Näiteks on mitmed 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 andmine, järelevalve, toetuse määramine) käigus saadud ja/või tekkivad andmed peaksid sisalduma ühes kirjeldatavas andmestikus, mitte olema pihustatud erinevate andmestike vahel. Näiteks ei ole mõistlik Ehitisregistri andmestiku piiritlemine selliselt, 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 võiks katta ühe või paar organisatsiooni funktsiooni. Näiteks on mõistlik defineerida andmestikuna “riigieelarve” või “keskkonnaseire andmed”, aga mitte “riigieelarve ja keskkonnaseire andmed”.

Peale (oluliste) andmestike defineerimist on mõistlik kaardistada detailsemalt ka andmestiku tehnilised komponendid - andmebaase, tabeleid, faile, andmeelemente jne. Kaardistuse eesmärgiks on eristada olulisemad ja vähemolulisemad komponendid, mis omakorda lubab edaspidi keskenduda ainult oluliste ja sisulist väärtust omavate andmete kirjeldamisele.

Näiteks on enamuses infosüsteemides talletatud küllalt palju tehnilist tugiinfot, mis on vajalik andmete turvalisuse ja terviklikkuse tagamiseks, kuid ei ole oluline andmehalduse ja -kirjelduse vaatest. Samuti võib andmestikust eksisteerida mitu paralleelset esitust nagu algne andmebaas ja selle põhjal loodud andmeait või andmeladu. Andmete kirjeldamiseks valmistumisel on tähtis erinevate tehniliste platvormide vahelistest seostest aru saada, analüüsida andmete kattuvusi ja taaskasutuse töövooge ning seeläbi tagada, et andmekirjeldusega on kaetud nii olulised andmed kui ka olulised platvormid (näiteks: hoone põhiandmete kirjeldus on seotud nii Ehitisregistri andmebaasi kui ka andmelao sobivate elementidega).

Andmestiku komponentide kaardistamisel tuleks tähele panna:

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

Selle etapi tulemusena on:

- ***Saadud kokkulepe, mis on kirjeldatav andmestik ja millised andmed selles sisalduvad***
- ***Saadud kindlus, et kirjeldamiseks valitud andmeid on mõttekas kirjeldada***
- ***Üles leitud kogu andmestiku kirjeldamist toetav dokumentatsioon***

3.2 Andmekirjelduse sõnastike kaardistamine

Andmekirjelduse koostamise teiseks sammuks on kirjeldamisel kasutatavas sõnavaras kokku leppimine, ehk valdkonna- ja ärisõnastike kaardistamine ning juba olemasoleva(te) andmekirjeldus(t)e tuvastamine.

Andmekirjelduse üks peamiseid 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 (mõisteid on tähistatud terminoloogiliselt üht moodi). Ühtse sõnavara kasutamine aitab organisatsioonil tõeliselt ära kasutada andmehalduse potentsiaali üle kõigi oma andmestike – 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 edasist kirjeldamist saab kiiremini ja efektiivsemalt teha, vältides vasturääkivusi kirjelduses ning hoides kirjeldamisel läbivalt ühtset sõnavara teiste allikatega.

Sõnastike kaardistamisel on esimeseks ülesandeks 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 ja pigem lähtuda selle sisu sobivusest. Olemasolevate rahvusvaheliste sõnastike ja ontoloogiate leidmiseks on nii mitmeid üldiseid katalooge,[10] kui ka avaliku sektori spetsiifikat kajastavaid sõnastikke ja ontoloogiad.[11]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 valdkonnaeksperdid võivad muude tegevuste käigus juba olla dokumenteerinud sobivaid sõnastikke. Lisaks organisatsiooni sees sõnastike kaardistamisele on võimalik abi küsida ka teistelt samas valdkonnas tegutsevatelt organisatsioonidelt, kuna mitmed valdkonnad kasutavad riiklikke või rahvusvahelisi valdkonnasõnastikke (nt. tervishoid, pangandus, looduskaitse).

Leitud sõnastike sobivuse hindamisel saab organisatsioon lähtuda järgmistest kriteeriumitest:

- **Ulatus ja kaetus**, ehk 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**, ehk 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,** ehk kui palju sarnaneb terminite kasutus sõnastikus õigusaktide ja/või organisatsiooni ja valdkonna ekspertide sõnavarale. Soovitatav on kasutada sõnastikke mis muude kriteeriumite

[10] Linked Open Vocabulary - https://lov.linkeddata.es/dataset/lov/; schema.org sõnastik - https://schema.org; Google'i sõnavara - https://developers.google.com/search/reference/overview; DBPedia projekti kaudu Wikipedia faktide esitamise sõnavara - https://wiki.dbpedia.org/services-resources/ontology

[11] EL - https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/core-vocabularies; BBC ontoloogiad - https://www.bbc.co.uk/ontologies

võrdsuse korral pakuvad paremat vastavust. Näiteks on ehitusseaduses defineeritud olulised terminid „hoone“ ja „rajatis“ kui üksteist välistavad: „rajatis on ehitis 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**, ehk kui suur on sõnastiku aktiivsete kasutajate hulk ja kui aktiivselt sõnastikku ennast hallatakse. Üldiselt on soovitatav eelistada laiemalt kasutatavaid sõnastikke, mille puhul on lisaks ka selgelt paika pandud sõnastiku haldamise ja edasiarendamise põhimõtted.

Praktikas leidub harva ühte kõikehõlmavat sõnastikku või ontoloogiat, mis kataks ära terve valdkonna eri tahud. Küll aga sisaldavad need enamasti üksikuid mõisteid, mis on andmekirjelduse jaoks piisavalt detailselt kirjeldatud. Seetõttu on loomulik kombineerida erinevaid sõnastikke ja ontoloogiaid vastavalt oma vajadusele.

Väljaspool organisatsiooni hallatava sõnastiku taaskasutamisel peaks organisatsioon eelkõige huvi tundma, kes on sõnastiku omanik ning kas ja millised protseduurid on paigas sõnastiku muutmiseks või täiendamiseks. Soovitatav on eelistada selliseid sõnastikke, mille omanik ja haldamise protseduurid on selgelt ja läbipaistvalt määratud. Juhul, kui organisatsioonil tekib vajadus teha ettepanek uute terminite lisamiseks või olemasolevate kirjelduste täiendamiseks, 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, siis on soovitatav paralleelsete sõnastike tekkimise ohu vältimiseks teha seda sama valdkonna organisatsioonide koostöös.

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

Juhul kui sobilikku sõnastikku ei leidunud, tuleb see n-ö nullist ise luua. Seda tegevust on kirjeldatud *Andmehalduse raamistiku* peatükkides 4.1.2 ja 4.1.3 ning käesoleva juhise peatükis 3.5.1.

Selle etapi tulemusena on:

- ***Tuvastatud andmestiku kirjeldamiseks sobivad organisatsiooni sisesed ja -välised sõnastikud***
- ***Analüüsitud, kas sõnastikud on kohased, piisavad, hästi hallatud ning tehniliselt kasutuseks sobivad***

3.3 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 eesmärgiks on kaardistada andmekirjelduse kasutusjuhud ja kasutajad nii organisatsiooni sees kui ka väljas. Sealjuures tuleks kasutusjuhte vaadelda laiemalt, nii et oleks kaasatud kõik olulised osapooled ja nende vajadused, sealhulgas:

- otsene andmekirjelduse koostamine ja sellega kokku puutuvad isikud (osapooled: andmehaldur, andmeomanik);

- andmebaasiplatvormis sisalduva info sünkroniseerimine andmekirjeldustega (andmebaasi administraator);
- andmekirjelduse taaskasutus suurandmete analüüsis või äriintelligentsuse (BI) rakendustes (andmeanalüütik);
- andmekirjelduse kasutamine andmekvaliteedi reeglite loomisel (andmehalduse juht, metaandmete analüütik);
- regulaarsete raportite koostamine juhtkonnale (juhtkond, 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). Näiteks tuleb X-teega ühendust eeldavate andmestike kirjeldused edastada RIHAsse, Statistikaamet kogub andmekirjeldusi riikliku statistika koostamiseks ja Rahvusarhiiv kogub andmekirjeldusi arhiiviväärtuslike andmete arhiveerimisel.

Lihtsamal juhul, näiteks paarist MS Exceli tabelist koosneva külastajate nimekirja puhul, võivad kasutusjuhud olla küllaltki lihtsad ning peamiseks vajaduseks on statistika edastamine juhile ja ministeeriumile. Keerulisemal juhul võib aga organisatsiooni andmekirjeldus katta mitmeid keeruka struktuuriga riiklike andmekogusid üle mitmete valdkondade ja sõnastike, ning seega võib ka kasutusjuhtude ja kasutajate tuvastamine olla keerukas ja aeganõudev.

Andmekirjelduse tehnilise arhitektuuri ülesandeks on kirjeldada tark- ja riistvaraline keskkond mis suudab mudelis tuvastatud vajadused piisavalt katta. Ülal toodud lihtsamal juhul on andmekirjelduse haldamine võimalik tavapäraste tabelarvutuse vahenditega (näiteks Microsoft Excel, OpenOffice ja LibreOffice Calc või Google Sheets), kus luuakse tabelid iga andmekirjelduse standardi (vt. Lisa 2) olemi jaoks: andmestiku ja andmeelementide kirjeldus ning andmesõnastik (vt. täpsemalt ptk. 3.5). Keerukamatel juhtudel on käsitsi andmekirjelduse tabelite koostamine, nendevaheliste seoste haldamine ning kirjelduse kasutajatele edastamine liialt aeganõudev, mistõttu on vaja juurutada spetsiaalne andmekirjelduse töövahend (vt. ka ptk. 2.5).

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

- Milliste andmeallikatega (näiteks andmebaasimootoritega) on andmekirjelduse töövahend võimeline (automaatselt) suhtlema;
- Millises vormingus sõnastikke ja kui lihtsalt on võimalik andmekirjelduse töövahendisse laadida;
- Kas andmekirjelduste koostamine on andmehaldurile või teistele kirjelduse loojatele mugav;
- Kas andmekirjelduse töövahend võimaldab andmekirjelduse standardi (vt Lisa 2) juurutamist;
- Kas ja millistes (avatud) vormingutes on andmekirjeldusi võimalik eksportida ning vajadusel välistele kasutajatele (Statistikaamet, RIHA, Rahvusarhiiv jt) edastada;
- Millises mahus on võimalik andmekataloogi tarkvara ja andmekirjelduse tegevusi integreerida teiste asutuse andmehalduse tegevustega (näiteks andmekvaliteedi haldus).

Selle etapi tulemusena on:

- ***Hinnatud täpselt andmekirjelduse mahtu ja vajadust***
- ***Analüüsitud sobilikke andmekirjelduse töövahendeid ja kirjelduse vorminguid***
- ***Valitud välja sobivaim lahendus ning selles loodud vajalik kirjelduse mudel ja struktuur***

3.4 Andmekirjeldusega seotud rollide määramine

Kui kirjeldatav andmestik ja olemasolevad sõnastikud on välja valitud, tuleb paika panna täpsem rollide jaotus andmekirjelduse loomiseks, haldamiseks ja kasutamiseks pikemas vaates. Praktikas on ärisõnastikul ja andmesõnastikul organisatsioonis sageli erinevad omanikud või vastutajad. Vajalik on nende omavaheline tihe koostöö sõnastike täiendamisel. Reeglina on neil erineva ulatusega juurdepääs andmetele ja muutmisõigused sõnastikele, milles nad ei ole andmeomaniku rollis. Praktikas sõltub õiguste halduse teostus oluliselt sõnastike haldamiseks kasutavast andmekirjelduse töövahendist. *Andmehalduse raamistiku* ptk. 3.3 kirjeldab põhjalikult andmehalduses osalevaid rolle ja nende ülesandeid. Andmekirjeldusega seotud rollide jaotus on järgmine:

Andmehalduse juht (juhtkonna liige) – korraldab ärisõnastike ja mõistete mudeli loomist juhtimise aspektist.

Andmeomanik (sisuteenuse juht või infosüsteemi peakasutaja) – loob ja kirjeldab teenuste käigus loodud või muudetud ärimõisted (ärisõnastik) ning nende seosed ärireeglitega organisatsiooni teistes infosüsteemides.

Andmehaldur (andmestike eest vastutav roll) – kirjeldab andmestikud ja andmeelemendid vastavalt andmekirjelduse standardile andmekirjelduse töövahendis.

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, kirjelduste koostamise eest ning näha ette järelevalvet teostav roll, mis jälgib andmekirjelduse kvaliteedi mõõdikute täitmist ning uuendab neid perioodiliselt (vt. ka ptk. 3.5.4).

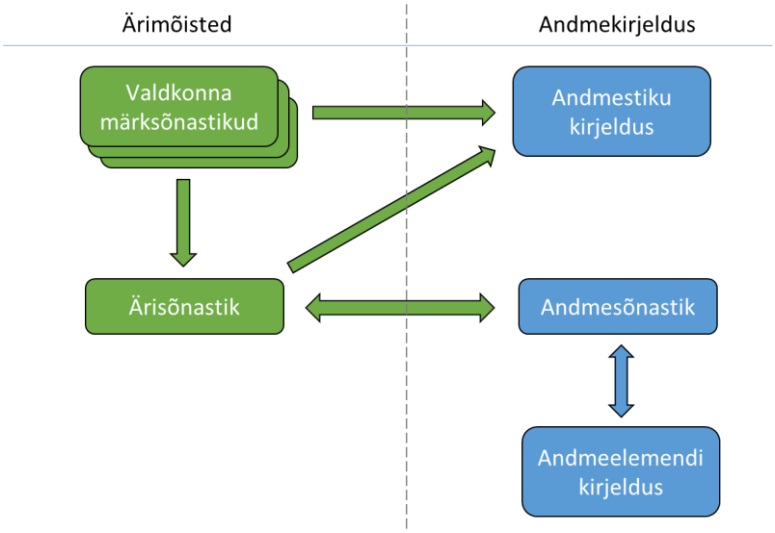
Rollide ja vastutuste jaotus on otstarbekas kirjalikult fikseerida organisatsiooni andmehaldust korraldavas dokumendis (näiteks „Andmehalduse reeglid“, vt. ka *Andmehalduse raamistik*, ptk. 6.7). Need rollid tuleb juurutada andmekirjeldusi ja sõnastikke haldavates tarkvararakendustes juurdepääsu- ja kasutusõigustena.

Selle etapi tulemusena on:

- ***Tuvastatud kõik organisatsioonis andmekirjeldusega seotud rolle täitvad isikud***
- ***Dokumenteeritud nende roll ja vastutus andmekirjeldusega seoses***

3.5 Andmekirjelduse loomine ja täiendamine andmekirjelduse töövahendis

Terviklik andmekirjeldus moodustub kolme kirjelduskomponendi sisestamisel ja omavahelisel seostamisel. Nendeks on: (1) andmestiku kirjeldus, (2) äri- ja andmesõnastikud ning (3) andmeelementide kirjeldus.

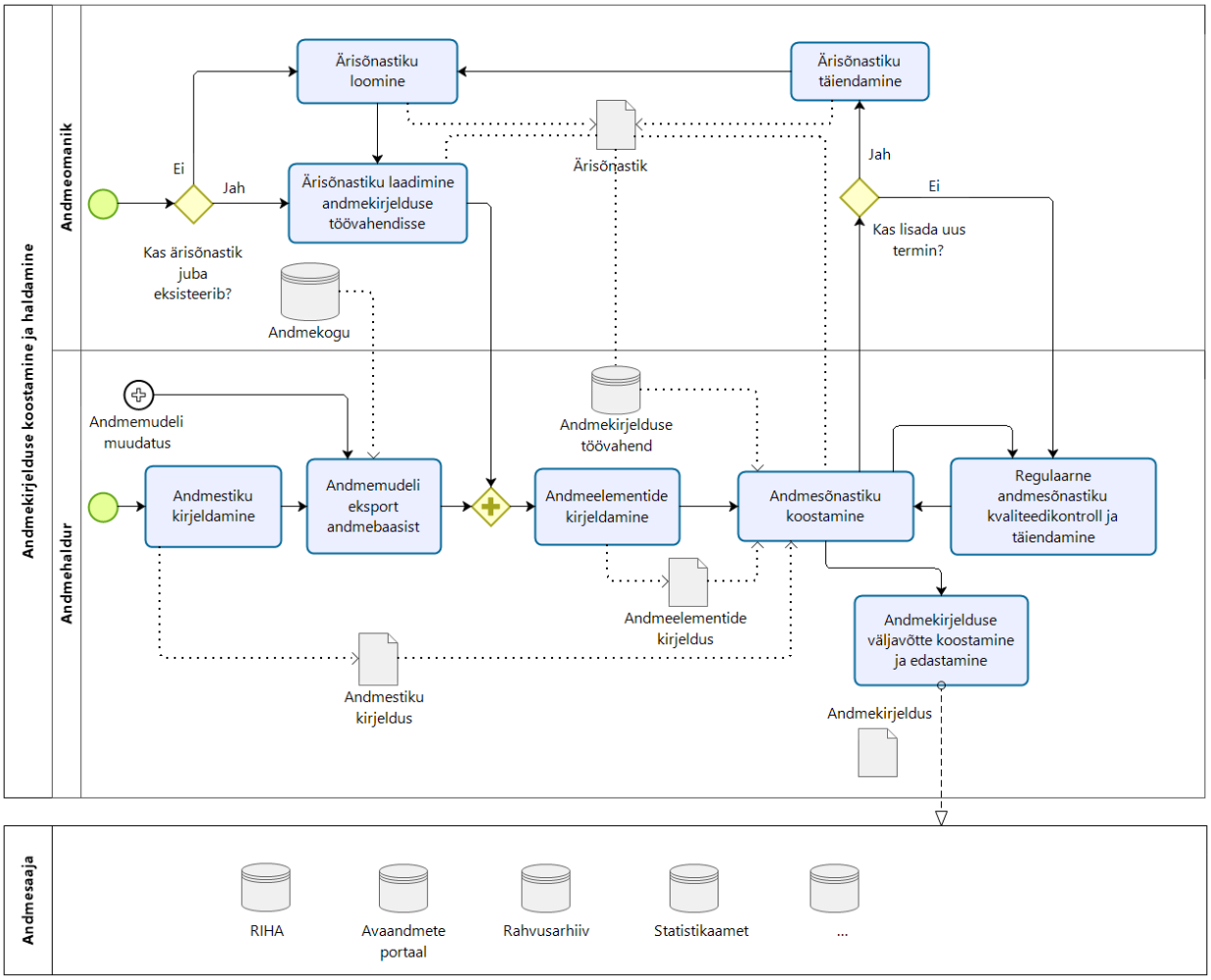


Joonis 6: Tervikliku andmekirjelduse komponendid

Joonis 7 allpool on esitatud andmekirjelduse loomise ja haldamise protsess, mida detailsemalt käsitlevad järgnevad alapeatükid (ptk. 3.5.1 – 3.5.4).

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 lahti seletavate terminitega.

Andmesõnastiku koostamine ei ole ühekordne hoogtöö, vaid pidev tegevus, kus nii erinevaid kirjeldusi kui ka kirjelduste vahelisi seoseid tuleb pidevalt hallata ja täiendada. Andmekirjelduse haldamine, parendamine ja erinevate komponentide vahel seoste loomine on seega pidev töö, mida tehakse järjepidevalt erinevate rollide koostöös.



Joonis 7: Andmekirjelduse loomine ja täiendamine andmekataloogis.

3.5.1 Ärisõnastiku loomine ja andmekirjelduse töövahendiga sidumine

Eeldus: *Ärisõnastiku loojal on hea ülevaade valdkonna terminoloogiast, sh olemasolevatest teistest seonduvatest ärisõnastikest, valdkonnasõnastikest, õigusaktides ja infosüsteemides kasutatud sõnavarast (vaata ka juhise peatükki* 3.2*).*

Organisatsiooni andmehalduse tegevusi toetava andmesõnastiku aluseks on läbivalt ühtlane 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, vaid 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 ristkasutada asutusesiseses või organisatsioonide vahelises andmevahetuses.
- Sõnastik on muutuv ja täienev. Valdkonna sõnavara, õigusaktide definitsioonid, aga ka hallatavate andmete skoop, on pidevas muutumises. Asutus peab nende muutuste toetamiseks looma kindla reeglistiku, mis kehtestab põhimõtted sõnastikku terminite lisamise, muutmise ja kustutamise kohta. Reeglite täitjaks ehk sõnastiku omanikuks, kes kogub kokku ettepanekud sõnastiku täiendamiseks ja kannab terminid sõnastikku, võiks üldjuhul olla andmehalduse juht või andmeomanik.

Ärisõnastiku loomisel on asutusel kolm peamist võimalust (vt. ka *Andmehalduse raamistiku* peatükke 4.1.2 ja 4.1.3):

- võetakse üle riiklik või rahvusvaheline valdkonna märksõnastik;
- asutus loob ärisõnastiku riikliku või rahvusvahelise märksõnastiku baasil;
- ärisõnastik luuakse asutuses oma tegevusvaldkonna õigusaktides ning andmestike ja infosüsteemide dokumentatsioonis kasutatud mõisteid tähistavate terminite baasil.

Need kolm võimalust ei ole üksteist välistavad. Näiteks võib asutus kasutada samaaegselt valdkonna märksõnastikku ja luua sellest puuduvate terminite jaoks õigusaktide definitsioonide kasutav asutuse ärisõnastik. 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 oleks kirjeldatud vastavalt andmekirjelduse standardis (vt. Lisa 2) esitatud nõuetele. Sõnastiku kirjelduse alusel peab olema võimalik vähemalt 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, ja kuidas on sõnastikku mõeldud kasutada. Seda võib teha kitsamalt ainult andmehalduse toetamiseks või laiemalt, näiteks asutuse ärireeglite, protsesside jms. kirjeldamiseks, linkandmete avaldamiseks, kommunikatsiooni parendamiseks jne. Iga mõiste juures peab sisalduma vähemalt termini tähenduse pikem kirjeldus, osutus kehtivusele (kas termin on aktuaalne või mitte) ja termini sõnastikku lisamise ja/või viimase muutmise kuupäev.

Ärisõnastik on oma ülesehituselt taksonoomiline ehk see esitab mõistete omavahelisi seoseid. Tavalised seose liigid on: hierarhiline („hoone“ ja „rajatis“ on mõlemad „ehitised“), antonüümia („hoone“ vs „rajatis“), sünonüümia („hoone“, „maja“) ning laiemaid ja kitsamaid termineid esitav kategoriaalne tesaurus („Euroopa“ riik on „Eesti“). Põhjalikumad sõnastikud võivad terminitega siduda ka organisatsiooni tööprotsesse, teenuseid ja neist tulenevaid ärireegleid ning seega esitada küllaltki keerulise mõistete ja seoste mudeli. *Andmehalduse raamistiku* ptk. 4.1.2 käsitleb mõistete mudeli visualiseerimise viise.

Ärisõnastiku tehnilise taaskasutuse toetamiseks on mõistlik igale terminile lisada masinloetav identifikaator (URI[12]) mille abil on võimalik iga terminit unikaalselt identifitseerida ja viidata. Muuhulgas saab identifikaatorit hästi kasutada ka terminite vaheliste seoste defineerimisel, kui URI struktuur selliselt luua. URIde automaatne loomine on toetatud ka mitmetes andmekirjelduse töövahendites.

Ärisõnastiku näide Ehitisregistri baasil lihtsa tabelarvutuse tabelina on toodud Tabel 1.

Tabel 1. Ärisõnastiku näide.

Mõiste	Selgitus	Ülemmõiste	Alam-mõiste	Infovara omanik	URI
Ehitis	Ehitis on hoone või rajatis.	Asi (kinnis- või vallasasi)	Hoone; Rajatis	MKM	http://ehr.ee/arisonastik/2020/ehitis
Hoone	Hoone on väliskeskkonnast katuse ja teiste välispiiretega eraldatud siseruumiga ehitis.	Ehitis		MKM	http://ehr.ee/arisonastik/2020/hoone
Rajatis	Rajatis on ehitis, mis ei ole hoone.	Ehitis		MKM	http://ehr.ee/arisonastik/2020/rajatis
Ehitis-registri kood	Ehitise kohta käivate andmete esmakordsel registreerimisel antakse ehitisele unikaalne ehitisregistri kood, milleks on numbrite kombinatsioon.	Ehitis		MKM	http://ehr.ee/arisonastik/2020/ehitiseEHRKood

[12] Uniform Resource Identifier https://www.ietf.org/rfc/rfc2396.txt

Ehitise nimetus	Ehitise nimetus annab edasi ehitise iseloomulikku tunnust, mille järgi on võimalik seda teistest eristada.	Ehitis		MKM	http://ehr.ee/arisonastik/2020/ehitiseNimetus
Kasutamise otstarve	Ehitise kasutamise otstarve vastavalt MKM 02.06.2015 määrusele nr 51 "Ehitise kasutamise otstarvete loetelu."	Ehitis		MKM	http://ehr.ee/arisonastik/2020/kasutamisotstarve

Ärisõnastiku loomiseks võib organisatsioon lihtsamal juhul kasutada tabelarvutustarkvara. Eelkõige on see mõistlik juhtudel, kui ka ülejäänud andmekirjelduse komponendid on plaanis luua lihtsate seotud tabelitena. Tuleb küll arvestada, et organisatsioonil on mõistlik ärisõnastik publitseerida (sise)veebis sellises vormingus, et see oleks andmekirjelduse koostamisel lihtsalt kasutatav märksõnaotsinguteks. Kui asutuses on kasutusel eraldi sõnastike haldamise või andmekirjelduse töövahend, tuleb ärisõnastik luua selliselt, et seda oleks võimalikult lihtne töövahendisse laadida. Käesoleval ajal on enim levinud ärisõnastiku loomine XML-põhiselt (näiteks OWL, RDFS, JSON-LD või SKOS skeemide alusel). Eestis on soovitatav OWL skeemi kasutamine, mida toetavad erinevad vabavaralised tööriistad, neist levinuim on Protege.[13]

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

Selle etapi tulemusena on:

- ***Välja valitud või koostatud ärisõnastik***
- ***Ärisõnastik seotud andmekataloogiga***

3.5.2 Andmekirjelduse loomine

Eeldus: *Andmekirjelduse looja on piiritlenud kirjeldatava andmestiku, tal on hea ülevaade andmete vormingust ja struktuurist, kirjeldatavad andmed on grupeeritud nende olulisuse alusel (vaata ka juhise peatükki* 3.1 Andmekirjelduse objektide piiritlemine*).*

Andmekirjeldus koosneb kahest peamisest komponendist: andmestiku kui terviku kirjeldus ja üksikute andmeelementide kirjeldus.

Andmestiku kirjelduse aluseks on Eestis DCAT standard,[14] mille nõutavad kirjelduselemendid on toodud andmekirjelduse standardis (Lisa 2). Kirjeldatakse andmestiku omanikku, eesmärke, kasutamist ja selles sisalduvate andmete põhitunnuseid. Täpsemad juhised kirjelduselementide kasutamiseks ja täitmiseks on leitavad standardist (Lisa 2).

Tabelis 2 on toodud andmestiku kirjelduse näide Ehitisregistri põhjal. Näites on võimalikult palju kirjeldusi täidetud erinevatest sõnastikest ja RIHAst tulenevate väärtustega. Ka teiste andmestike kirjeldamisel on soovitatav võimalikult suures mahus juba olemasolevaid allikaid taaskasutada.

[13] https://protege.stanford.edu/

[14] https://www.w3.org/TR/vocab-dcat-2/

Tabel 2. Andmestiku kirjelduse näide Ehitisregistri baasil.

Nr	Kirjelduselement	Väärtus	Selgitus
1	Andmestiku nimi	Ehitisregister	Andmestiku nimetus õigusaktis või asutuses kasutatav nimetus
2	Andmestiku kirjeldus	Ehitisregistri eesmärgiks on Eesti Vabariigis asuvate ehitiste ning nendega seotud menetluste kohta andmete koondamine, hoidmine ja avalikkusele juurdepääsu võimaldamine. Andmekogus on andmed ehitatavate ja kasutatavate ehitiste, ehitusgeoloogiliste ja -geodeetiliste tööde ja nende teostajate, ehitusprojektide ja ehitiste ekspertiiside, vallasasjast ehitiste omanike, ehitus- ja mõõdistusprojektide koostajate ja kontrollijate, ehitamist teostavate ja teostanud isikute, ehitus-projektide ja ehitiste ekspertiiside ning omanikujärelevalve teostajate, vallasasjast ehitisega seotud pantide, arestide ja keeldude ning ehitise ja ehitamisega seotud ettekirjutuste ning ehitamisega seonduvate taotluste, teatiste ja lubade kohta.	Andmestiku pidamise eesmärgi ja andmete sisuline lühikirjeldus
3	Andmestiku identifikaator	70003158-ehitisregister	Soovitatav on identifikaatori loomisel kasutada järgmist struktuuri: [asutuse registrikood]-[andmestiku lühinimi]
4	Andmestiku õiguslik alus	Ehitisregistri põhimäärus https://www.riigiteataja.ee/akt/128062019019	Viide andmestiku loomise ja haldamise aluseks olevale õigusaktile
5	Andmestiku teema	tööstus/ehitus	Kontrollitud märksõnastikust võetud märksõna, siin näide EuroVoc taksonoomia alusel
6	Andmestiku märksõna	ehitis; energiamärgis; ehitised	Organisatsiooni ärisõnastikust, valdkondlikust või üldisest märksõnastikust võetud üks või mitu märksõna. Näite allikad: Ehitisregistri ärisõnastik, Eesti üldine märksõnastik
7	Andmestiku märksõna URI	http://ehr.ee/arisonastik/2020/ehitis http://ehr.ee/arisonastik/2020/energiamargis https://ems.elnet.ee/id/EMS016903	Märksõna viide URI vormingus
8	Andmestiku tüüp	Dataset	Kirjeldatava andmestiku tüüp loendist Dublin Core Type Vocabulary (DCMI Type vocabulary)
9	Andmestiku keel	et	Andmestikus kasutatud keel(ed)
10	Andmestiku omanik	Majandus- ja kommunikatsiooniministeerium	Andmestiku omanik-organisatsiooni (nt. andmekogu vastutav töötleja) täielik nimetus
11	Andmestiku haldaja	Majandus- ja kommunikatsiooniministeerium	Andmestiku haldaja-organisatsiooni (nt. andmekogu volitatud töötleja) täielik nimetus
12	Andmestiku kontakt	N, N, ehr@mkm.ee, 625 6363	Andmestiku eest vastutava töötaja kontaktandmed
13	Andmestiku veebisait	https://www.ehr.ee/ https://opendata.riik.ee/andmehulgad/ehitisregister/	Andmestiku avalik veebiliides või muu veebisait millelt on võimalik andmestikku kasutada või sellele juurdepääsu taotleda

14	Andmestikuga seotud organisatsioon	kohalikud omavalitsused, Tarbijakaitse ja Tehnilise Järelevalve Amet, Muinsuskaitseamet, Raudteeamet, notarid, energiamärgise väljastajad	Andmestikuga seotud organisatsioon(id), mis kas esitavad andmestikku andmeid, või kasutavad andmestiku andmeid
15	Andmestiku juurdepääsupiirangud	Piiratud juurdepääs taotlustele, teatistele ja ettekirjutusele lisatud füüsiliste isikute andmetele (AK, AvTS §35 lg 1 punkt 12; §40 lg 3)	Kogu andmestikule kohalduvate kasutustingimuste kirjeldus
16	Andmestiku kasutuslitsents	CC BY-SA	Kogu andmestikule kohalduv Creative Commons kasutuslitsents
17	Andmestiku seos	lubade väljastamine; tingimuste väljastamine; teatiste väljastamine; ehitusjärelevalve	Organisatsiooni teenuste loetelust võetud teenuste nimetused, mille osutamiseks selle andmestiku andmeid kasutatakse
18	Andmestiku päritolu	(eelkäija) Riiklik Hooneregister (eelkäija) Riiklik Ehitusregister	Teise andmekogu nimetus, millest andmestiku andmed on päritud
19	Andmestiku standard	Building & Land Development Specification (BLDS)	Loetelu rahvusvahelistest või valdkondlikest standarditest, mida andmestik kasutab
20	Andmestiku alguskuupäev	2003-01-01	Andmestiku pidamise algusaeg ehk andmestiku infosüsteemi loomise või kasutuselevõtmise aeg
21	Andmestiku muutmiskuupäev	2020-05-28	Andmestiku viimase muutmise kuupäev
22	Andmete piirdaatumid	1992 -	Andmestikus sisalduvate andmete piirdaatumid
23	Andmete uuendamise regulaarsus	Pidev	Andmestiku andmete uuendamise regulaarsus
24	Andmete ruumiline ulatus	Eesti	Haldusüksuse nimetus ja tüüp, mida andmestiku andmed katavad

Ühe andmestiku kirjeldus ei ole mahukas, aga mitmed kirjelduselemendid (näiteks kirjeldus, seosed, märksõnad) eeldavad põhjalikku ja läbimõeldud sisu, mida mõnel juhul tuleb käia otsimas organisatsiooni veebisaidilt, Riigi Teatajast või ka küsida andmeomaniku käest.

Andmeelementide kirjeldamisel tuleb Eestis kasutada kohandatud ja lokaliseeritud DDI standardit,[15] mille nõutavad kirjelduselemendid on toodud andmekirjelduse standardis (vt. Lisa 2). Kirjeldada tuleb üksikuid andmebaasi veerge nii tehniliste kirjeldustega (näiteks veeru nimi, andmetüüp), mis on leitavad andmebaasi andmemudelist, kui sisestada andmevälja inimloetav nimi, lühikirjeldus ja kasutusele võtmise aeg. Vajadusel peab viitama ka välja väärtuse aluseks olevale klassifikaatorile või loendile. Lisaks andmeelementidele peab üldiselt kirjeldama ka andmestiku füüsilist struktuuri ehk väljade vahelisi seoseid (võtmeid). Täpsemad juhised kirjelduselementide kasutamiseks ja täitmiseks on leitavad standardist.

Andmeelemendi kirjeldus Ehitisregistri näitel on toodud tabelis 3. Paksus kirjas on toodud kirjelduselemendid mida on võimalik automaatselt lugeda andmebaasist.

Tabel 3. Andmeelemendi kirjeldus Ehitisregistri näitel.

Nr	Kirjelduselement	Väärtus
1	**Andmeelemendi tähis**	KOETAV_PIND
2	Andmeelemendi GUID	26d728e7-7eee-4f6a-a7ba-a0a62831d947
3	Andmeelemendi URI	http://ehr.ee/EHR/EH_EHITIS/KOETAV_PIND
4	Andmeelemendi nimetus	Köetav pind
5	**Andmeelemendi kirjeldus**	Ehitise köetav kogupind (m^2)
6	Andmeelemendi märksõna URI	http://ehr.ee/andmesonastik/2019/koetavPind
7	**Andmeelemendi tüüp**	number(16)
8	Seotud loendi tähis	-
9	Andmeelemendi väärtuse mõõtühik	ruutmeeter (m^2)
10	Andmeelemendi väärtuse kordaja	-
11	Andmeelemendi kasutamise algus	2003-01-01
12	Andmeelemendi kasutamise lõpp	2019-01-21
13	Viide primaarvõtmele	-
14	**Andmeelemendi tabel**	EH_EHITIS

Kui lihtsamaid andmestikke võib kirjeldada käsitsi, siis suuremad andmebaasid võivad koosneda sadadest tabelitest ja tuhandetest andmeelementidest. 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 seda 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.

Andmekirjelduse koostamise viimaseks sammuks on andmesõnastiku loomine. Oma olemuselt on andmesõnastik andmeelementide nimetuste ja ärisõnastiku ühenduslüli, ehk ta seob omavahel ärisõnastiku mõiste ning sellele vastavad andmeelemendid. Kuigi andmesõnastik on sarnaselt ärisõnastikule terminipõhine ja paljud andmesõnastiku terminid kattuvad ärisõnastiku omadega, on neil kolm olulist erinevust mis tingivad andmesõnastiku vajaduse eraldi komponendina:

[15] https://ddialliance.org/Specification/DDI-Lifecycle/3.3/

- **Omanik:** ärisõnastiku omanikuks ja haldajaks on andmete omanik, andmesõnastiku omanikuks on andmete kirjeldaja (andmehaldur). Seega on andmete kirjeldajal lähtuvalt andmekirjelduse vajadustest võimalik andmesõnastikku termineid kiirelt ise luua, samas kui ärisõnastikku termini lisamine võib olla küllaltki pikaajaline ja keeruline protsess;
- **Ulatus:** andmesõnastikus kajastatakse ainult andmeelementide kirjeldamisel vajalikud mõisted. Seega ei pruugi selles sisalduda kõik ärisõnastiku terminid, samuti võib andmesõnastikus sisalduda termineid mis ei eksisteeri ärisõnastikus.
- **Keerukus:** ärisõnastiku terminid võivad olla omavahel seotud küllaltki keerulise mudeli alusel, andmesõnastik koosneb peamiselt omavahel seostamata terminitest ja sõnaseletustest.

Andmesõnastiku terminid saadakse kolmel viisil.

- Need võivad olla ärisõnastikus olevad terminid, kui need mõisteliselt kattuvad kirjeldatava andmeelemendi tähendusega.
- See võib olla andmeelemendi nimetus.
- See võib olla andmesõnastikku loodud termin, mis on saadud mitme andmeelemendi rühmitamisel üheks mõisteks ja selle tähistamisel terminiga, mida ärisõnastikus ei olnud. Selle termini võib hiljem ka ärisõnastikku kanda (vt. ptk. 3.5.3).

Andmesõnastiku koostamist on võimalik alustada nii andmeelementide kui ärisõnastiku vaatest:

- Alustades andmeelementidest on andmesõnastiku koostaja tegevuseks üksikute andmeelementide rühmitamine (näiteks „eesnimi“, „perekonnanimi“, „isikukood“), sellele andmeelementide grupile uue andmesõnastiku termini loomine (näiteks „isik“) või sobiva ärisõnastikus juba olemasoleva terminiga seostamine (näiteks „isik“).
- Alustades ärisõnastikust toimub sama tegevus ülalt alla – esmalt valitakse ärisõnastikust termin, kasutatakse seda andmesõnastiku terminina 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 seonduvaid andmeelementide komplekte võib leiduda mitmes erinevas asukohas (tabelis). Sellisel juhul ei tohi iga andmeelementide komplekti jaoks luua uut sõnastiku terminit, vaid siduda **kõik** andmeelemendid ühe terminiga. Andmesõnastiku termini kirjelduse näide Ehitisregistri baasil on toodud Tabel 4.

Tabel 4. Andmesõnastiku termini kirjelduse näide.

Nr	Kirjelduselement	Väärtus
1	Termin	Köetav pind
2	Termini määratlus	Ruumide pind, mille õhu temperatuur reageerib kütteperioodil välisõhu temperatuuri muutustele vaid vähesel määral. Arvestatakse ruutmeetrites.
3	Termini URI	http://ehr.ee/andmesonastik/2019/koetavPind
4	Termini staatus	Kinnitamata
5	Termini kehtivus	Jah
6	Termini allikas	Majandus- ja taristuministri määrus Nõuded energiamärgise andmisele ja energiamärgisele, § 2 (14)
7	Termini loomiskuupäev	2003-01-01
8	Termini muutmiskuupäev	2019-01-29

9	Termini seosed	http://ehr.ee/arisonastik/2019/koetavPind; http://riha.eesti.ee/riha/onto/ettevotlus.toostussektorid.ehitised/2020/r1/koetavPind
10	Märkused	Kuni 21.01.2019 kehtis määratlus „Hoone köetav pind on hoone kõigi sisekliima tagamisega ruumide suletud netopindade summa.“, kuid seoses energiamärgise jaoks arvestatava köetava pinna erisustega asendus mõiste määratlus. Ehitise köetava pinna hilisem väärtus erineb madalatemperatuuriliste köetavate pindade erineva arvestuse võrra. Kui energiamärgisele on vaja mõistet, mis ei sisaldaks madalatemperatuurilisi köetavaid pindu, on vaja selleks luua uus ning selgelt eristuv mõiste, mille väärtusi salvestatakse andmestikus uuele väljale.

Andmesõnastiku koostamisel on üheks nõudeks kõigi kirjelduse komponentide, milleks on 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 hierarhilisteks komponentideks on identifitseeritav objekt, näiteks http://www.asutus.ee/sonastik/termin. Samas ei pea andmete kirjeldaja ise selliste identifikaatorite loomisega tegelema, need tekitatakse automaatselt andmekirjelduse töövahendi poolt. Eestis kokku lepitud andmekirjelduse standardis on samuti nõutud kõigi andmekirjelduse komponentide identifitseerimine URI skeemi alusel.

Andmekirjelduse standard sisaldab ka andmesõnastikku kirjeldavaid metaandmeid. Need loovad sõnastikule vajaliku konteksti juhuks, kui seda on vaja avaldada või jagada teiste organisatsioonidega. Sõnastiku metaandmed on talletatud reeglina andmekirjelduse töövahendis.

Tabel 5. Andmesõnastiku kirjelduse näide.

Nr	Kirjelduselement	Väärtus
1	Andmesõnastiku nimi	Ehitisregistri andmesõnastik
2	Andmesõnastiku kirjeldus	Ehitisregistri andmesõnastik sisaldab ehitisregistri andmeobjektide kirjeldust
3	Andmesõnastiku URI	http://ehr.ee/ehitised/dataDictionary/v2/
4	Andmesõnastiku omanik	Majandus- ja Kommunikatsiooniministeerium
5	Andmesõnastiku eelmine versioon	http://ehr.ee/ehitised/dataDictionary/v1/
6	Kasutab sõnastikke	http://ehr.ee/ehitised/BuildingsOntology/v3/
7	Andmesõnastiku loomise kuupäev	2020-09-23
8	Andmesõnastiku viimase muutmise kuupäev	2020-10-19

Selle etapi tulemusena on:

- ***Valminud põhjalik ja läbimõeldud andmestiku kirjeldus***
- ***Kirjeldatud piisav hulk andmeelemente***
- ***Loodud seosed andmeelementide, andmesõnastiku ja ärisõnastiku vahel***

3.5.3 Andme- ja ärisõnastiku täiendamine

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

Juba andmesõnastiku esialgse koostamise käigus võib ilmneda, et kõigi oluliste andmeelementide jaoks ei ole ärisõnastikus leida 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 alustamise tõttu, mis omakorda toob kaasa vajaduse andme- ja ärisõnastiku täiendamiseks. Andmesõnastiku kaudu lisandub sedasi ärisõnastikku uusi termineid või

võib täpsustamist vajada mõiste seletus (vt. näidet Tabel 4. Andmesõnastiku termini kirjelduse näide. elemendis 10 Märkused). Samuti võib ärisõnastik uueneda tänu uutele ülesannetele, mis organisatsioon on saanud või võtnud ning organisatsioon võib otsustada võtta ärisõnastikuna kasutusele mõni uus valdkonna märksõnastik.

Üldjuhul on andmesõnastiku ja ärisõnastiku omaniku rollid erinevate töötajate kanda, mistõttu eeldab sõnastike uuendamine tihedat omavahelist suhtlust ja selgelt paika pandud protseduure (vt. ka *Andmehalduse raamistik*, ptk. 4.1.3). Algselt on üsna tavaline, et ühte ja sama on nimetatud erinevate terminitega. Selliste terminite vaheliste konfliktide lahendamisel on üldise reeglina ärisõnastikus olev termin ülimuslik andmesõnastiku termini ees. Andmesõnastike uuendamise tüüpilised stsenaariumid ja nendega seotud otsused on:

- Andmeelementide kirjeldamisel ja rühmitamisel luuakse andmesõnastikku termin, kuid sellele ei leita sobivat vastet ärisõnastikust:
  1) Andmesõnastiku omanik teeb ettepaneku uue termini ehk märksõna lisamiseks ärisõnastikku, põhjendades termini vajadust, viidates nii andmeelementidele kui termini selgituse allikale. Kuni kinnitamiseni on andmesõnastiku termini staatus „kinnitamisel“.
  2) Ärisõnastiku omanik kaalub märksõna lisamist ja kinnitab või lükkab ettepaneku tagasi. Otsuse tegemine võib olla kollektiivne, nt. ärisõnastiku haldamise töörühmas.
  3) Kinnitamisel lisatakse märksõna kinnitatud staatuses ärisõnastikku ja lisatakse andmesõnastikku viide märksõnale. Andmesõnastikus saab termini staatuseks „kinnitatud“.
  4) Sõnastiku täiendamise tagasilükkamine peab olema põhjendatud, nt. soovitus kasutada mõnd teist ärisõnastiku märksõna. Negatiivse otsuse kohta lisatakse teade märksõna juurde ärisõnastikus, andmesõnastikus tuleb teha kas soovituse kohane muudatus või jätta termin kasutusse vaid andmesõnastikus koos staatusega „kinnitamata“.
  5) Uuendatud ärisõnastik seostatakse või laetakse andmekirjelduse töövahendisse.
- Organisatsiooni tegevuse, teenuste või tööprotsesside uuenemise tulemusena võetakse ärisõnastikus kasutusele uus märksõna või muudetakse olemasoleva seletust:
  1) Ärisõnastiku omanik kontrollib, kas uuenenud märksõnad on seotud andmeelementidega ja teavitab uuendustest andmesõnastiku omanikku.
  2) Andmesõnastiku omanik analüüsib muudatuste mõju ja sobivust andmesõnastikule ja teeb otsused kas:
     - kinnitada uus termin andmesõnastikku lisandusena ja siduda sellega uued (loodavad) andmeelemendid;
     - kinnitada olemasolevate termini uus seletus andmesõnastikku;
     - lisada andmesõnastikku uus termin ja jagada andmesõnastikus olemasoleva terminiga seotud andmeelemendid olemasoleva ja uue termini vahel.
     - jätta uus termin andmesõnastikku lisamata.
  3) Uuendatud ärisõnastik seostatakse või laetakse andmekirjelduse töövahendisse.

Selle etapi tulemusena on:

- ***Defineeritud ärisõnastiku ja andmesõnastiku täiendamise protseduurid***

3.5.4 Andmekirjelduse kvaliteedikontroll

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

Andmekirjelduse kvaliteedi peamiseks mõõdikuks on probleemideta taaskasutus, ehk kasutajate võimekus andmekirjeldustest aru saada. Andmekirjelduse kvaliteedile hinnangu andmisel saab lähtuda kahest peamisest küsimusest:

1) 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.
2) Kas kirjeldus on piisavalt kvaliteetne, et tagada andmete ühene mõistetavus nii organisatsiooni siseste kui väliste kasutajate jaoks? Sellele küsimusele vastamiseks saab organisatsioon mõõta näiteks:
    - Andmekirjelduse kohta organisatsioonile esitatud probleemide hulka võrdluses andmekirjelduse kasutajate arvuga;
    - Andmekirjelduse põhjal loodud andmekvaliteedi reeglite hulka.

Oluline on välja tuua, et andmekirjelduse üks peamiseid kasutusstsenaariumeid ongi viimases näites mainitud teenuse ärireeglitest tulenevate 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 taaskasutamist ja edastamist käsitleb järgmine alapeatükk, aga ka selle toetamine on üks andmekirjelduse haldamise pidevatest toimingutest. Andmekirjelduse haldamist on otstarbekas käsitleda pideva parendamise tsüklina – planeeri, teosta, kontrolli, korrigeeri (PDCA), kus eesmärgiks on andmekirjelduse kvaliteedi tõstmise mõõdikud.

Andmehalduri tüüpilised tegevused andmekirjelduse kvaliteedi tagamisel on:

- Regulaarne andmesõnastiku ja andmeelementide kirjelduse ülevaatus. See võib toimuda kas pisteliselt või/ja üksnes hiljuti muudetud ja lisatud kirjelduste osas.
- Valdkonnaspetsialistide intervjueerimine andmekirjelduse ja -sõnastiku kasutatavuse ja ühese mõistetavuse osas.
- Andmete organisatsiooni siseste kui väliste kasutajate toetamine ja neilt tagasiside kogumine.

Selle etapi lõpuks on:

- ***Defineeritud protseduurid andmekirjelduse regulaarseks ülevaatuseks***
- ***Korraldatud andmekirjelduse kvaliteedi hindamine ja vastava tagasiside kogumine kasutajatelt***

3.6 Andmekirjelduse edastamine

Andmekirjelduste tähtsaks kasutusjuhuks on organisatsiooni sisene või organisatsioonide vaheline andmete taaskasutus. Organisatsiooni siseselt võib andmekirjelduste edastamine olla vajalik näiteks andmeintegratsiooni, virtualiseerimise või andmelao keskkonda, uute andmeteenuste loomisel, andmejagamisplatvormi või andmepõhise aruandluse jaoks automatiseeritud näidikukaardi lahenduse loomisel jmt.

Organisatsioonide vaheliselt on andmekirjelduste vahetamine tihti vajalik sama valdkonna asutuste vahel. Samuti nõuavad andmekirjelduste edastamist õigusaktid, näiteks RIHA (avaliku sektori andmetest tervikülevaate haldamisel) või riigi ülesannete täitmine, näiteks Statistikaamet riikliku statistika kogumisel ja Rahvusarhiiv arhiiviväärtuslike andmete kogumisel riigi digitaalarhiivi. Samuti võib asutusel olla vaja avalikustada andmed koos nende kirjeldusega avaandmetena, edastades andmed riigi avaandmete portaali.[16] Andmekirjelduse kasutajaks on üha enam teine tarkvara. See tõstab andmekirjelduse masinmõistetavuse tähtsust, mille hulgas väga oluline mõistete ühene identifitseeritavus läbi seda kasutatavate URIde, nimeruumidele viitamine ja ontoloogiate kasutamine.

Asutuste või protsesside vahelist andmete jagamist realiseeritakse läbi mõistete avalike sõnastike (vt. ka *Andmehalduse raamistik*, ptk. 4.1.10). Andmehalduse raamistikus ette nähtud andmevahetus-kokkulepped ja -nõuded kasutavad nii ärisõnastikku kui andmekirjeldusi. Üldarusaadavate mõistete ühitamine eri andmestike või teenuste vahel võimaldab vältida ohtu, et vahetama asutakse üleliigseid või ebausaldusväärseid (nt. vananenud) andmeid. Andmevahetuse toetamiseks peab andmekirjeldus olema ajakohane ja piisav, et olla üheselt mõistetav ka väljaspool valdkonnaspetsialistide kitsast ringi.

Kindlasti on igal andmekirjelduse kasutajal oma eripärad, vajadused ja nõuded. Enamasti on kirjelduste kasutaja huvitatud ainult osast kirjeldusest andmestiku mingi alamhulga kohta, mis tuleb eraldi väljavõttena luua või defineerida.

Iga andmekirjelduse edastamise juhtumi jaoks on vaja:

- Teha andmekirjelduse saaja vajaduste ja toetatud vormingute ning liideste analüüs.
- Defineerida kogu andmekirjeldusest saajale sobilik alamosa ja esitada see väljavõttena, sidudes see teenuse osutamiseks andmestiku vastava alamosaga või seadistada vastav teenust osutav liides.
- Viia andmekirjelduse väljavõte saajale sobivasse struktuuri ja vormingusse (näiteks XML teisenduse abil).
- Edastada andmekirjelduse väljavõte 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 kasutamine on enamasti siiski regulaarne või pidev, näiteks asutuse protsess uute andmekirjelduste edastamiseks andmelattu kord nädalas või kirjelduste edastamine RIHAsse iga andmeandmekirjelduses 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. Näiteks Statistikaametile hoonete andmekirjelduse väljavõtte tegemise ja edastamise profiil. Andmekirjelduse edastamise vajadused (liidesed, kirjelduse vormingud jms) võivad seetõttu oluliselt mõjutada organisatsiooni otsuseid andmekirjelduse töövahendivalikul.

[16] https://opendata.riik.ee/

3.7 Andmekirjelduse seostamine organisatsiooni tööprotsesside ja teenustega

Andmehalduse raamistik (vt. ptk. 4.1.3, 4.1.5) paneb paika reeglid, kuidas andmeelementidega seotud andmekvaliteedi nõuded seotakse organisatsiooni tööprotsesside ja teenustega. Selliste reeglite kaudu on võimalik tagada, et andmete kvaliteet vastab tööprotsesside ja teenuste kvaliteedi nõuetele. Igale mõistega kirjeldatud andmeelemendile määrab andmeomanik andmekvaliteedi reegli, mis kirjeldab, millistes piirides on tulevased andmed kvaliteetsed. Näiteks lihtne nõue, et mingi kindel andmeväli ei või kunagi jääda tühjaks, nagu kliendi aadress, hüvitise summa, vmt. Selliste nõuete olemasolu võib oluliselt tõsta andmete kasutatavust ja kasulikkust organisatsioonis (vt. ka *Andmekvaliteedi juhis*, ptk. 3.2). Sõnastikel ja andmekirjeldusel on teenuste kvaliteedi tagamisel seega oluline roll sillana andmete ja ärireeglite vahel.

Organisatsiooni tegevust kajastavatest peamistest mõistetest moodustuv ärisõnastik on seotud andmestikku kirjeldava andmesõnastikuga. Nii ärisõnastikku, andmeelementide kirjeldusi kui ka moodustuvat andmesõnastikku tuleb pidevalt võrrelda sellega, milliste terminoloogiaga on kirjeldatud organisatsiooni tööprotsesse ja teenuseid. Protsesside ja teenuste muutumisel tuleb täiendada nii sõnastikke kui ka kirjeldusi ning andmekvaliteedi reegleid. See on pidev töö andmekirjelduse kvaliteedi tõstmiseks (vt. ptk. 3.5.4 eespool) ja andmekvaliteedi reeglite rakendamiseks. Selle saavutamiseks peavad ühelt poolt eksisteerima toimivad andmehalduse protsessid ning teiselt poolt protsessijuhtide ja teenuseomanike strateegiline huvi. Andmehaldurite töö mõistete ühtlustamisel nii organisatsiooni sõnastike piires, andmestike kaupa kui ka organisatsioonide vahel, tõstab andmekvaliteeti, muudab andmete tähenduse paremini mõistetavaks ja soodustab andmete taaskasutust (vt. ptk. 3.6 eespool).

Sõnastike halduse ja andmekirjelduse töövahendite kasutamine on üks võimalus tagada tõhus kontroll asutuse sõnavara, metaandmete ja andmete kvaliteedireeglite üle. Ärisõnastike haldamise vahendeid on kirjeldatud *Andmehalduse raamistiku* peatükis 5.2 ja neid on võimalik juurutada nii laiemate platvormide osana (nt. Atlassian tooteperekond) kui ka iseseisvate rakendustena (nt. Informatica, Collibra, Erwin). Andmekirjelduse töövahenditest on lühidalt juttu käesoleva juhise peatükis 2.5 ja Lisas 1.

Lisa 1: Andmekirjelduse koostamise ja haldamise abivahendid

Andmekirjeldust reguleerivad õigusaktid

Avaliku sektori kõige olulisemaks õiguslikult reguleeritud osaks on andmekogud. Need on loodud seadusega või seaduse alusel määrusega ning neil on põhimäärus. Andmekogu põhimääruse sisu kirjeldab avaliku teabe seaduse § 43[5] lõige 1: „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 eri andmekogude puhul loetletud erineva üksikasjalisusega. Mõnel juhul on esitatud andmete üldised rühmad nagu üldandmed, teisel juhul andmeobjektid ja kolmandal juhul detailsed andmeelementidena käsitletava andmed. Näiteks Töökeskkonna andmekogu põhimääruses on toodud loetelu faktidest, mille kohta infot kogutakse ja salvestatakse. Haigekassa andmekogu põhimääruses on toodud andmeväljade loend, mida andmekogus hoitakse. Töövõime hindamise ja töövõimetoetuse andmekogus on toodud objektide loend ja iga objekti kohta väljade loend (lühikese kirjeldusega), mis andmekogus salvestatakse.

Ühtlustatud kujul on andmekogude kohta esitatavate andmekirjelduse elementide loetelu toodud Vabariigi Valitsuse määruses riigi infosüsteemi haldussüsteem.[17] Selle § 18 „RIHA andmekogude alamregister“ loetleb lõikes 2 alamregistrisse kantavate andmekogu üldandmete koosseisu, milles on 27 kirjelduselementi; lõige 3 andmekogus töödeldavate andmete koosseisu, milles on 6 kirjelduselementi ning lõige 4 andmekogu asutamise ja andmete töötlemise aluseks olevate õigusaktide andmete koosseis, milles on 4 kirjelduselementi. Nendega on käesolevas juhises arvestatud.

Riigi infosüsteemi koosvõime puhul eristatakse õiguslikku, organisatsioonilist, tehnoloogilist ja semantilist koosvõimet.[18] Andmekirjelduse kontekstis, mille eesmärgiks on saavutada parem andmetest arusaamine ja kvaliteet, on oluline eelkõige semantiline koosvõime, mis tähendab erinevate organisatsioonide võimet mõista vahetatavat andmeid ja informatsiooni ühtmoodi. Riigi semantilise koosvõime raamistikku uuendati viimati aastal 2007,[19] mistõttu see ei kajasta vahepeal toimunud arenguid semantikavahendites ning paraku ei ole andmekirjeldamise korraldamisel abiks.

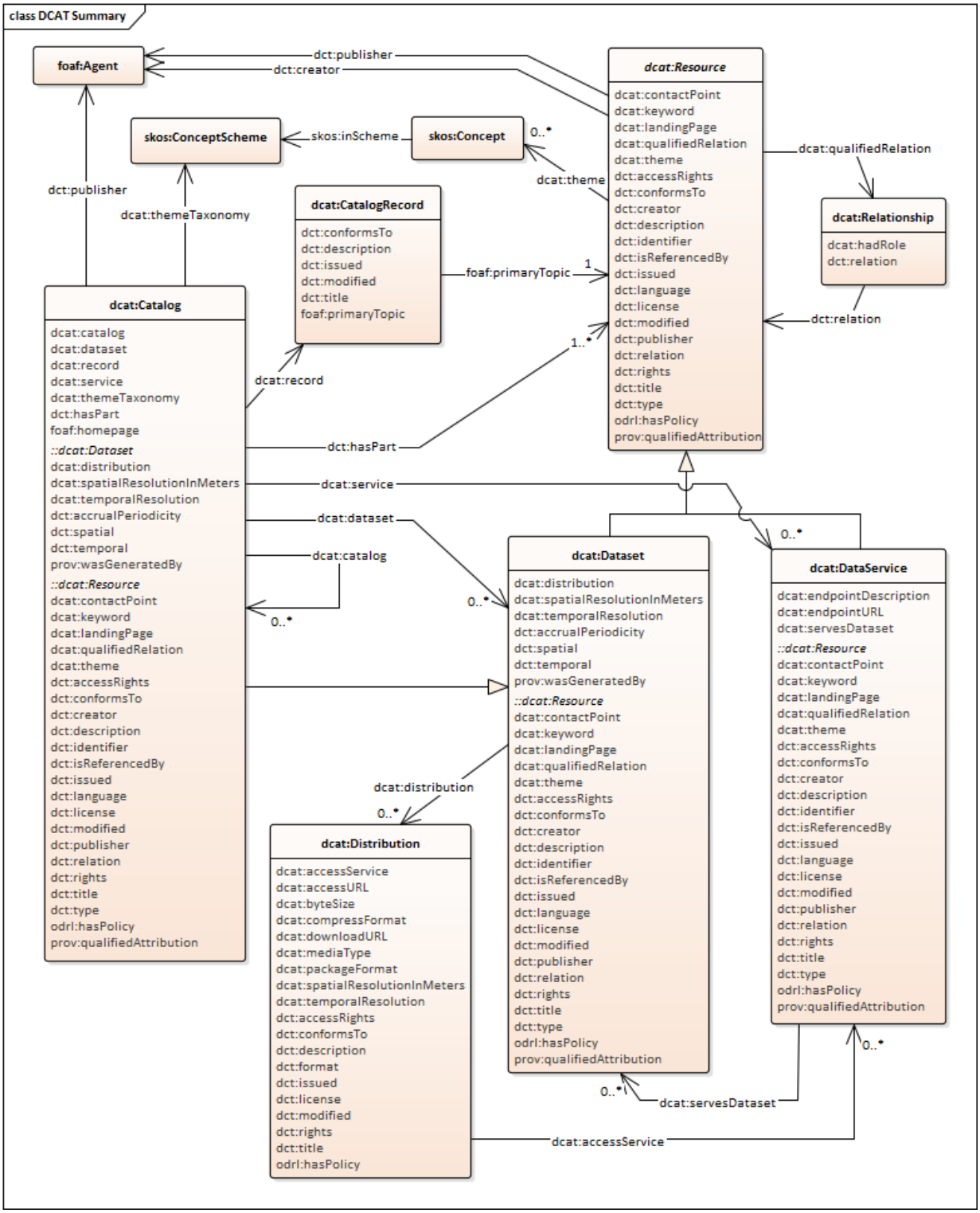
Andmekirjeldusega on seotud ka Vabariigi Valitsuse määrus Klassifikaatorite süsteem,[20] mis määratleb klassifikaatorite haldamise ja kasutamise ühtsed põhimõtted.

Andmekirjeldusega seotud standardid

DCAT – Data Catalog Vocabulary (Andmekataloogi sõnastik), versioon 2 (2020)
https://www.w3.org/TR/vocab-dcat-2/

DCAT on oma olemuselt RDF sõnastik, mis on loodud selleks, et aidata kaasa veebis avaldatud andmekataloogide koostalitusvõimele. DCAT-i kasutamine andmestike kirjeldamiseks tagab standardse mudeli ja sõnastiku rakendamise, mis aitab kaasa eri andmekataloogides olevate metaandmete kasutamisele ja üldistamisele ning parandab andmestike ülesleitavust. Kasutades DCAT-i on võimalik andmekataloogi detsentraliseeritud avaldamine ning otsingu teostamine üle mitme kataloogi, kasutades sama päringu mehhanismi ja struktuuri. Rühmitatud DCAT metaandmed on kasutatavad sisukirjeldusena, mis on osaks andmete säilitusprotsessist.

[17] https://www.riigiteataja.ee/akt/106082019018
[18] https://www.mkm.ee/et/tegevused-eesmargid/infouhiskond/riigi-infosusteem
[19] https://www.mkm.ee/sites/default/files/riigi_infosusteemide_semantilise_koosvoime_raamistik.pdf
[20] https://www.riigiteataja.ee/akt/12910889



Joonis 8. DCAT mudeli klassid.

DDI - Data Documentation Initiative Lifecycle (Andmete dokumenteerimise algatus), versioon 3.3 (2020) https://ddialliance.org/Specification/DDI-Lifecycle/3.3/

DDI on rahvusvaheline kirjeldusstandard statistiliste uuringute ja vaatluste, küsimustike, statistiliste andmefailide ja sotsiaalteadusliku uuringu tasemel info haldamiseks. DDI metaandmed on esitatud XML skeemi kujul.

ISO/IEC 11179 Information technology — Metadata registries (MDR) (Metaandmete registrid) on rahvusvaheline standard, mis käsitleb andmete semantikat, andmete kirjeldamist ja esitamist nende taaskasutamiseks ning andmekirjelduse registreerimist. Standard koosneb kokku kaheksast osast: raamistik, liigitamine, registri metamudel, andmemääratluste koostamine, nimetamise põhimõtted, registreerimine, andmestiku kirjelduse metamudel ning metaandmete põhiomadused. Standardi käsitluses on metaandmeelementidel mitu olekut: 1) lihtsalt identifitseeritud ühes andmekataloogis kasutamiseks; 2) identifitseeritud ja registreeritud, ehk hallatud keskses metaandmete registris (nt. andmekataloogis). Metaandmete registri pidamise reeglistik on standardi üks osa ja puudutab organisatsioonilist korraldust.

GSIM – Generic Statistical Information Model (Üldine statistiline info mudel), versioon 1.2 (2019) https://statswiki.unece.org/display/gsim/GSIM+v1.2+documents

SDMX – Statistical Data and Metadata eXchange (Statistikaandmete ja metaandmete edastamine), versioon 2.1 (2013) https://sdmx.org/?page_id=5008

Seotud juhised

Andmekirjelduse juhisega paralleelselt valmisid 2020. aasta kevadel *Andmehalduse raamistik*,[21] mis käsitleb andmehalduse terviklikku juurutamist organisatsioonis ja *Andmekvaliteedi juhis*,[22] mis kirjeldab andmekvaliteedi mõõtmist ja andmekvaliteedi jaoks reeglite kehtestamist, ning *Andmehalduse teadmusbaas*,[23] mis sisaldavad täiendavalt viiteid kasulikele juhendmaterjalidele.

RIA ja PwC koostöös välja töötatud *Andmekvaliteedi tagamise juhend andmekogu omanikele* (2016)[24] esitab andmekvaliteedi juhtimise raamistiku ja andmekvaliteedi haldamise protsesside kirjelduse, s.h. mitmeid andmekirjeldusega seotud teemasid (vt. andmete kooskõla, andmete usaldusväärsus, jt.).

Andmemudelite semantilise kirjeldamise käsiraamat *SEHKE* (2009)[25] juhendab semantikavarade kasutamist andmete kirjeldamisel.

RIA *Avaandmete loomise ja avaldamise juhend* (2016)[26] käsitleb andmete semantilist mõistetavust.

Arhiiviväärtuslikuks hinnatud andmekogude üleandmist Rahvusarhiivile kirjeldab rahvusarhiivi juhis *Arhivaalide üleandmine*.[27] Rahvusarhiiv arhiveerib ainult arhiiviväärtuslikuks hinnatud andmekogusid,[28] millel eeldatakse andmekirjelduse olemasolu ja selle üleandmist arhiveerimisel.[29]

[21] http://www. … .ee/
[22] http://www. … .ee/
[23] http://www.digiriik.ee/
[24] https://www.ria.ee/sites/default/files/content-editors/publikatsioonid/andmekvaliteedi_tagamise_juhend_andmekogu_omanikele.pdf
[25] http://lambda.ee/wiki/Sehke_speci_draft
[26] ria.ee/sites/default/files/content-editors/publikatsioonid/avaandmete_loomise_juhend.pdf
[27] http://www.ra.ee/arhiivihaldus/juhised/#6
[28] http://www.ra.ee/wp-content/uploads/2016/11/LISA-1-HO-nr-51-31.10.2017-ANDMEKOGUD.pdf
[29] http://www.ra.ee/arhiivihaldus/digitaalarhiivindus/andmekogude-arhiveerimine/

Teiste riikide andmekirjelduse materjale

Andmekirjelduse juhendeid ja nõudeid on välja antud enamuses Euroopa riikides, kuid valdavalt vaid kohalikus riigikeeles. Siin alapeatükis toome mõned näited materjalidest, mis on kättesaadavad ingliskeelsena.

Ameerika Ühendriikides on kasutusele võetud DCAT standardi lokaliseeritud versioon,[30] mis on mõeldud eelkõike rakendamiseks avaandmete kirjeldamisel ja avaldamisel.

Austraalias on kehtestatud üleriigiline valitsussektori funktsioonide tesaurus (AGIFT),[31] kus on kolmetasemelisse hierarhiasse seatud nii keskvalitsuse kui kohalike omavalitsuste ülesanded.

Austraalia Rahvusarhiiv on avaldanud koosvõimet toetavate metaandmete juhendi,[32] mis aitab organisatsioonidel välja töötada metaandmete strateegia, rakendada metaandmete haldusvahendeid ning hallata metaandmete hoidlat. Muu hulgas sisaldab see ka andmekirjelduse rolli illustreerivaid andmevahetuse stsenaariume.[33]

Uus-Lõuna Wales'i provintsi andmehalduse tööriistakast sisaldab näpunäiteid andmekirjelduse edukaks rakendamiseks.[34]

METeOR[35] on Austraalia rahvuslik meditsiini, tervishoiu ja hoolekande valdkonna metaandmete hoidla. See on suurepärane näide valdkonna märksõnastike, ärisõnastike ja andmekirjelduse elementide terviklikust ja kontseptualiseeritud haldamisest.

Uus-Meremaa riiklik andmehalduse raamistik[36] on terviklik käsitlus andmehalduse erinevatest aspektidest, mis muu hulgas hõlmab ka andmesõnastiku loomist.[37]

Andmekirjelduse töövahendid

Üldised märksõnastikud

ELNET Konsortsium haldab *Eesti Märksõnastikku* https://ems.elnet.ee/index.php
Eesti Keele Instituudi mitmekeelne terminibaas *Estterm* https://termin.eki.ee/esterm/

Valdkondlike märksõnastike näiteid

Andmeanalüüsi ja statistika oskussõnastik https://term.eki.ee/termbase/view/8917007/

Raamatukogusõnastik https://termin.nlib.ee/

EstCore2 https://projektid.hitsa.ee/display/HAK/EstCORE+2

Andmekirjelduse tööriistad

Äri- ja andmesõnastike haldamise töövahendid on välja kasvanud metaandmete hoidla rakendustest (*metadata repository*), mis koondasid kokku organisatsiooni andmelaos kasutatavad mõisted ja

[30] DCAT-US Schema v1.1 (2014) https://resources.data.gov/schemas/dcat-us/v1.1/
[31] Australian Governments' Interactive Functions Thesaurus https://data.naa.gov.au/def/agift.html
[32] https://www.naa.gov.au/information-management/building-interoperability/interoperability-development-phases/data-governance-and-management/metadata-interoperability
[33] https://www.naa.gov.au/sites/default/files/2019-09/Interoperability%20scenarios.pdf
[34] https://data.nsw.gov.au/data-governance-toolkit
[35] https://meteor.aihw.gov.au/content/index.phtml/itemId/181414
[36] https://www.data.govt.nz/manage-data/data-stewardship/
[37] https://www.data.govt.nz/manage-data/data-stewardship/creating-a-data-dictionary/

andmeelementide metaandmed ning võimaldasid neid (pool)käsitsi hallata. Praeguseks on andmekataloogi (*data catalog*) tarkvarad arenenud ja mõeldud nii ärisõnastiku loomiseks kui andmesõnastiku pidamiseks ning äri- ja andmesõnastiku vaheliste seoste haldamiseks, mille kõrval võib olla ka võimalusi andmeelementidele kehtestatud andmekvaliteedi reeglite kirjeldamiseks. Andmekataloogi lahenduste funktsionaalsused võivad oluliselt erineda – alates andmeobjektide kirjeldusest ja inventuurist kuni andmeteadlasele vajalike tööriistadeni - või on need erinevalt jaotatud tarkvara moodulite vahel. Samuti võib erineda suutlikkus toetada erinevaid andmekirjelduse mudeleid ja liidestada valdkondlikke märksõnastikke.

Tarkvaraturul on andmekirjelduse ja andmekataloogi toodete valik lai, nii integreerituna andmetöötluse platvormidega või eraldiseisvate toodetena. Gartner avaldab regulaarselt ülevaadet juhtivatest toodetest.

Joonis 9. Andmekataloogi tarkvarade võrdlus. Allikas: *Gartner 2019*[38]

Vabavaraliste toodete hulgas on põhjalikuma andmekirjelduse funktsionaalsusega näiteks Egeria raamistik,[39] TrueDat[40] ja Magda.[41]

Käesoleva juhise koostamisega samaaegselt viisid Riigi Infosüsteemide Amet ja Statistikaamet koostöös Majandus- ja Kommunikatsiooniministeeriumiga läbi analüüsi ja prototüübi arenduse projekti

[38] G. De Simoni, M. Beyer, A. Jain, *Gartner Magic Quadrant for Metadata Management Solutions* (2019)
[39] https://egeria.odpi.org/
[40] https://www.truedat.io/product/
[41] https://magda.io/

andmekirjelduse haldamise vahendi loomiseks asutustele, projektinimega RIHAKE. Selles kasutatav andmekirjelduse mudel tugineb käesolevas juhises toodud andmekirjelduse standardile (vt. Lisa 2). Kava kohaselt täidaks RIHAKE asutuses andmekirjelduse töövahendi rolli ja selles saab hallata nii andmesõnastikku kui ka andmestiku ja andmeelementide kirjeldusi ning selles on olemas funktsionaalsus asutuses loodud andmekirjelduste edastamiseks RIHAsse ja teistele seotud osapooltele.

Lisa 2: Andmekirjelduse standard

Kirjeldatavad olemid on järjestatud tähestikulises järjekorras.

Andmeelemendi kirjeldus

#	Elemendi nimetus	Määratlus ja kasutamine	Kohustuslik	Näide	Standardi viide	RDF element
1	Andmeelemendi tähis	Tehniline tähis, mis võib olla täheline, numbriline või muu lühend või akronüüm. Tähis võib olla semantiliselt arusaadav, kuid ei pruugi seda olla.	Jah	algus_kpv; haridus; eluk_EHAK; jt28	DDI 3.3:[42] element: VariableName > attribute:r:String	
2	Andmeelemendi GUID	Andmeelemendi globaalselt unikaalne identifikaator.	Jah	GUID: 123e4567-e89b-12d3-a456-426655440000	DDI 3.3: element: Variable > element:r:ID	
3	Andmeelemendi URI	Nimest, aadressist või tähisest koosnev URI, mis viitab andmeelemendile ainuliselt. Märkus: URI võetakse andmesõnastike jaoks kasutusele koos RIHAKEse juurutamisega.	Jah	http://ehr.ee/ ehitisregister/v2.0/schema:EHR2012/ table:Omanik/element:OmanikEesnimi	DDI 3.3: element: Variable > r:URN	
4	Andmeelemendi nimetus	Kas määratlus (ehk definitsioon), pealkiri (mis järgib pealkirjastamise reegleid) või terminoloogiline tähistus (märksõna), mis avab elemendi mõistelise sisu.	Jah	Küsitluse alguse kuupäev; Haridustase; Püsielukoht EHAK-tasemel	DDI 3.3: element: Variable > r:Label	
5	Andmeelemendi kirjeldus	Lühike kirjeldus andmeelemendi allikast, kontekstuaalsest tähendusest, kasutusest jms.	Jah	Andmeelemendi „Püsielukoht EHAK-tasemel“ kirjeldus: Püsielukoht on elukoht, kus isik veedab enamiku oma igapäevasest puhke- ja uneajast. Püsielukoht loetakse olevat Eestis, kui isik on pidevalt elanud vähemalt 12 kuud enne küsitlust Eestis.	DDI 3.3: element: Variable > r:Description	

[42] Andmeelemendi kirjelduses viidatakse Data Description Initiative standardi versiooni 3.3 (2020) elementidele, vt. https://ddialliance.org/Specification/DDI-Lifecycle/3.3/

				EHAK on Eesti haldus- ja asustusjaotuse klassifikaator.		
6	Andmeelemendi märksõna URI	Viide andmesõnastiku terminile / sõnale juhul kui andmeelement vastab andmesõnastiku terminile või on osaks sellest. Märkus: URI võetakse andmesõnastike jaoks kasutusele koos RIHAKEse juurutamisega.	Ei	http://ehr.ee/dataDictionary/v2/term:Ehitise_brutopind	DDI 3.3: element: Keyword > attribute: controlledVocabularyURN	
7	Andmeelemendi tüüp	Andmeelemendi tüüp andmebaasist, koosneb enamasti põhitüübist (text, numeric, date jne) ning pikkusest või vormingust (text[60], varchar[2]). Kui andmeelement on seotud koodiloendiga (klassifikaatoriga), siis peab olema väljas *Seotud koodiloendi tähis* (vt. element nr. 8) märgitud, mis koodiloendiga on tegu.	Jah	varchar[80]; date; datetime	DDI 3.3: element: Variable > VariableRepresentation	
8	Seotud loendi tähis	Koodiloendi või klassifikaatori andmekogu siseselt või üldiselt kokku lepitud lühend/akronüüm. Kirjeldus on kohustuslik kui elemendi väärtus järgib loendit või klassifikaatorit (vt. element nr. 7).	Jah*	ER_Ravikuuritüüp; EHAK; AK2008ap; taskOutcomeSearch	DDI 3.3: element: CodeListName > attribute:r:String	
9	Andmeelemendi väärtuse mõõtühik	Andmeelemendi sisu mõõtmiseks kasutatav ühik. Kirjeldus on kohustuslik, kui element nr. 7 Andmeelemendi tüüp eeldab väärtuste mõõtühiku ära näitamist.	Jah*	sekund (s); meeter (m); kilogramm (kg)	DDI 3.3 element: MeasurementUnit	
10	Andmeelemendi väärtuse kordaja	Kirjeldus on kohustuslik, kui element nr. 7 Andmeelemendi tüüp eeldab väärtuste mõõtühiku ära näitamist ning kordaja on kasutusel.	Jah*	Tuhandetes tonnides; miljonit eurot		

11	Andmeelemendi kasutamise algus	Kuupäev, mis ajast hakati elementi andmetega väärtustama ehk andmeid selle elemendi kohta täitma (vormingus AAAA-KK-PP).	Ei	2010-01-01		
12	Andmeelemendi kasutamise lõpp	Kuupäev, mis ajast elementi enam ei täideta, mille all ei mõelda mitte juhuslikku täitmist või mittetäitmist, vaid otsust elementi mitte täita (vormingus AAAA-KK-PP).	Ei	2020-04-18		
13	Viide primaarvõtmele	Kui andmeelement on väline võti (*foreign key*), esitatakse siin viide primaarvõtmeks olevale andmeelemendile. Kirjeldamisel kasutatakse vormingut [TABEL].[VÄLJA TÄHIS] .	Ei	Viide tabeli KORTER välja HOONE_ID kirjelduses: HOONE.ID; Viide tabeli JARELEVALVE välja TEOSTAJA_ID kirjelduses: ISIK.ID		
14	Andmeelemendi tabel	Tabeli, millesse andmeelement kuulub, tähis.	Jah	HOONE; ISIK		

* Kirjelduselement on tingimuslikult kohustuslik, ehk kohustuslik juhul kui andmeelemendis on vastav väärtus (klassifikaator, mõõtühik, väärtuse kordaja) rakendatud

Andmestiku kirjeldus

#	Elemendi nimetus	Määratlus ja kasutamine	Kohustuslik	Näide	Standardi viide	RDF element
1	Andmestiku nimi	Andmekogu korral selle pidamist reguleerivas õigusaktis toodud ametlik nimetus. Muu andmestiku puhul võimalusel õigusaktis toodud täielik nimetus või praktikas kasutatav täielik nimetus.	Jah	Ehitisregister; Asutuse N külastuste register	DCAT 6.4.6[43]	dct:title
2	Andmestiku kirjeldus	Andmestiku pidamise eesmärgi ja andmete sisuline lühikirjeldus. Andmekogu korral selle asutamise õigusaktis toodud eesmärk koos sisu kirjeldusega; muude andmestike korral loomise kirjeldus.	Jah	Keskkonnaregister on loodusvarade, looduspärandi, keskkonnaseisundi ja keskkonnategurite andmeid sisaldav riigi põhiregister. Keskkonnaregistri eesmärk on koondada kogu keskkonnaandmestik ühte registrisse, seostades selle kaudu kõik keskkonnaandmed ajas ja ruumis ning anda neile õiguslik tähendus, tagades sellega andmestiku usaldatavuse nii rahvusvahelisel kui siseriiklikul tasandil.	DCAT 6.4.5	dct:description
3	Andmestiku identifikaator	Andmestiku lühinimi. Soovitatav on identifikaatori loomisel kasutada järgmist struktuuri: [asutuse registrikood]-[andmestiku lühinimi] .	Jah	70000332-ametipalk (Statistikaameti andmestik http://andmestikud.stat.ee/ametipalk/); 75018710-haudi (Muhu valla kalmistute andmekogu)	DCAT 6.4.11	dct:identifier
4	Andmestiku õiguslik alus	Viide andmestiku loomise ja haldamise aluseks olevale õigusaktile (nt. andmekogu	Ei	Valga valla jäätmevaldajate registri põhimäärus	DCAT 6.4.22	dct:isReferencedBy

[43] Andmestiku kirjeldamisel viidatakse Data Catalog Vocabulary versiooni 2 (2020) elementidele, vt. https://www.w3.org/TR/vocab-dcat-2/

		põhimäärus). Võimalusel kasutada nii õigusakti nimetust kui URLi (Riigi Teatajas, organisatsiooni kodulehel).		https://www.riigiteataja.ee/akt/407062018011		
5	Andmestiku teema	Kontrollitud märksõnastikust võetud märksõna. Märkus: hetkel ei ole sobivat märksõnastikku kokku lepitud, kuid tegu on avaliku sektori funktsioonide klassifikaatorile (COFOG) lähedase hierarhilise klassifikaatoriga. Valitakse üks klassifikaatori kõige alumise taseme märksõna.	Jah	[haridus/algharidus]	DCAT 6.4.12	dcat:theme
6	Andmestiku märksõna	Organisatsiooni ärisõnastikust, valdkondlikust või üldisest märksõnastikust võetud üks või mitu märksõna. Eelistada tuleb üldisi asutuse tegevusi kirjeldavaid termineid.	Ei	jäätmekäitlus; liikluskorraldus	DCAT 6.4.16	dcat:keyword
7	Andmestiku märksõna URI	Märksõna viide URI vormingus. Kasutatakse koos eelmise elemendiga juhul kui märksõna allikaks olev sõnastiku termin on viidatav.	Ei	http://asutus.ee/arisonastik/2020/termin https://ems.elnet.ee/id/EMS016903	-	-
8	Andmestiku tüüp	Kirjeldatava andmestiku tüüp. Loendist Dublin Core Type Vocabulary võetud üks väärtus. Vaikimisi väärtuseks on "dataset" ehk andmestik, võimalusel saab kasutada ka täpsemat väärtust nagu näiteks "text", "picture".	Ei	dataset	DCAT 6.4.13	dct:type
9	Andmestiku keel	Andmestikus kasutatud keel(ed). Kui kirjeldus pole täidetud, eeldatakse vaikimisi eesti keele kasutamist.	Ei	et	DCAT 6.4.9	dct:language
10	Andmestiku omanik	Andmestiku omanik-organisatsiooni täielik nimetus. Andmekogu kirjeldamisel märgitakse siin vastutava töötleja nimetus, muude andmestike korral selle	Jah	Majandus- ja kommunikatsiooniministeerium; Kihnu vald	DCAT 6.4.4	dct:creator

		organisatsiooni nimetus, kelle valitsemise all tegelikult andmed on.				
11	Andmestiku haldaja	Organisatsiooni täielik nimetus. Andmekogu kirjeldamisel märgitakse siin andmekogu volitatud töötleja nimetus, muude andmestike korral märgitakse andmestiku omanik.	Jah	RMIT; Audru vald	DCAT 6.4.10	dct:Publisher
12	Andmestiku kontakt	Töötaja nimi, amet, e-post ja telefon, kes on andmestiku eest vastutav. Soovitav kasutada organisatsiooni andmehalduri kontakte.	Ei	Jüri Tamm, andmehaldur, jüri.tamm@mmit.ee, 6789012	DCAT 6.4.3	dcat:contactPoint
13	Andmestiku veebisait	Andmestiku avalik veebiliides või muu veebisait millelt on võimalik andmestikku kasutada või sellele juurdepääsu taotleda. Ühe andmestiku kohta võib kirjeldada mitu veebisaiti, näiteks 1) avalik portaal, 2) juurdepääsutaotluste esitamist kirjeldav / võimaldav veebileht, 3) asutuse avaandmete portaal.	Ei	http://register.keskkonnainfo.ee/envreg/; https://haridus.saaremaavald.ee/huviharidus/	DCAT 6.4.17	dcat:landingPage
14	Andmestikuga seotud organisatsioon	Andmestikuga seotud organisatsioon(id). Võimaldab kirjeldada asutusi või asutuste rühmi mis kas a) esitavad andmestikku andmeid, või b) kasutavad andmestiku andmeid. Kirjeldamisel sisestatakse organisatsiooni täielik nimetus või rühma üheselt mõistetav nimetus.	Ei	Audru vald; notarid; KOVid	DCAT 6.4.18	prov:qualifiedAttribution
15	Andmestiku juurdepääsu-piirangud	Kogu andmestikule kohalduvate kasutustingimuste kirjeldus. Näiteks on sobilikud tingimus: avalik andmestik; metaandmed on avalikud,	Ei	Andmestiku metaandmed on avalikud, andmed on mitteavalikud	DCAT 6.4.1	dct:accessRights

		andmed osaliselt avalikud; metaandmed avalikud, andmed mitteavalikud; metaandmed osaliselt avalikud, andmed mitteavalikud; vmt.				
16	Andmestiku kasutuslitsents	Kogu andmestikule kohalduv kasutuslitsents, kui see on määratud. Kasutada Creative Commons litsentse.	Ei	CC-By	DCAT 6.4.19	dct:license
17	Andmestiku seos teenusega	Organisatsiooni teenuste loetelust võetud teenuste nimetused loeteluna, mille osutamiseks selle andmestiku andmeid kasutatakse.	Ei	Lemmiklooma registreerimine; Hauaplatsi hooldaja määramine	DCAT 6.4.15	dcat:qualifiedRelation
18	Andmestiku päritolu	Teise andmekogu nimetus, millest andmestiku andmed on päritud. Võimalik on kirjeldada kahe tüüpi päritolu: a) andmestiku eelkäija, või b) andmestiku aluseks olev teine andmestik. Võimalusel tuleks elemendis kasutada teise andmekogu unikaalset identifikaatorit.	Ei	(eelkäija) Hooneregistri infosüsteem; (eelkäija) 70001975-skais; (alusandmestik, rahvastikuregister) 70000562-rr	-	prov:wasDerivedFrom > prov:wasRevisionOf (eelkäija) prov:hadPrimarySource (alusandmestik)
19	Andmestiku standard	Loetelu rahvusvahelistest või valdkondlikest standarditest, mida andmestik kasutab. Mõeldud on andmestikku tervikuna puudutavat standardit (nt. meditsiinivaldkonnas SNOMED või aruandluses XBRL).	Ei	BLDS	DCAT 6.4.2	dct:conformsTo
20	Andmestiku alguskuupäev	Andmestiku pidamise algusaeg ehk andmestiku infosüsteemi loomise või kasutuselevõtmise aeg. Vt. ka element nr. 23 Andmete piirdaatumid, mis käsitleb ka võimalikke varasemaid andmete tekkimise aegu, mis on andmestikku üle kantud.	Jah	2003-01-10	DCAT 6.4.7	dct:issued

21	Andmestiku muutmiskuupäev	Andmestiku viimase muutmise kuupäev. Viimase muutmise kuupäev täidetakse ainult suletud või perioodiliselt uuenevate andmestike puhul.	Ei	2020-05-25	DCAT 6.4.8	dct:modified
22	Andmete piirdaatumid	Andmestikus sisalduvate andmete piirdaatumid, näitab millist ajaperioodi andmed katavad. Andmete piirdaatumeid ei tohi segi ajada andmekogumise või -sisestamise piirdaatumitega. Näiteks 1920 aasta paberdokumendi andmete sisestamisel on kuupäevaks 1920, mitte 2020. Aktiivselt täiendatava andmestiku puhul pole lõppdaatumit tarvilik märkida.	Ei	1993-11-05; 2007-06-30; 1920; 1940	DCAT 6.6.5	dct:temporal
23	Andmete uuendamise regulaarsus	Andmestiku andmete uuendamise regulaarsus. Täita juhul kui andmete kogumine või loomine ei toimu pidevalt vaid perioodiliselt.	Ei	Kord aastas	DCAT 6.6.6	dcat:temporalResolution
24	Andmete ruumiline ulatus	Haldusüksuse nimetus ja tüüp. Eesti kohta kasutatakse jaotust vald/linn, maakond või kogu Eesti. Muu maailma kohta riigi nimetus, regiooni nimetus või globaalne.	Ei	Eesti; Mulgi vald; Põhjamaad; Soome	DCAT 6.6.3	dct:spatial

Andmesõnastiku kirjeldus

#	Elemendi nimetus	Määratlus ja kasutamine	Kohustuslik	Näide	OWL/RDF element[44]
1	Andmesõnastiku nimi	Sõnastiku nimetus, üldjuhul koos viitega sõnastiku omanik organisatsioonile.	Jah	Kultuuriministeeriumi andmesõnastik	rdfs:label
2	Andmesõnastiku kirjeldus	Andmesõnastiku sõnaline kirjeldus, mis viitab selle kasutusvaldkonnale või ulatusele.	Ei	Spordi-, loomemajanduse- ja etenduskunstide valdkondade andmestike andmesõnastik	rdfs:comment
3	Andmesõnastiku URI	Viide andmesõnastiku nimeruumile (URI). Märkus: URI võetakse andmesõnastike jaoks kasutusele koos RIHAKEse juurutamisega.	Jah	http://ehr.ee/ehitised/dataDictionary/v2/	rdf:about
4	Andmesõnastiku omanik	Andmesõnastiku omanik-organisatsiooni nimi.	Ei	Majandus- ja Kommunikatsiooniministeerium	dct:creator
5	Andmesõnastiku eelmine versioon	Viide andmesõnastiku eelmisele versioonile (URI). Kasutatakse juhul, kui eelmine versioon on olemas. Märkus: URI võetakse andmesõnastike jaoks kasutusele koos RIHAKEse juurutamisega.	Ei	http://ehr.ee/ehitised/dataDictionary/v1/	owl:priorVersion
6	Kasutab sõnastikke	Viited andmesõnastikus kasutatavatele teistele sõnastikele, nt. asutuse ärisõnastik (URI). Märkus: URI võetakse andmesõnastike jaoks kasutusele koos RIHAKEse juurutamisega.	Ei	http://ehr.ee/ehitised/BuildingsOntology/v3/	owl:imports
7	Andmesõnastiku loomise kuupäev	Andmesõnastiku kasutusse võtmise kuupäev.	Jah	2020-09-23	dct:created
8	Andmesõnastiku viimase muutmise kuupäev	Andmesõnastiku viimase täiendamise kuupäev.	Ei	2020-10-19	dct:modified

[44] Andmesõnastiku kirjeldamisel kasutatakse võimalusel W3C Web Ontology Language (OWL, https://www.w3.org/OWL/) standardis sõnastiku kirjeldamiseks defineeritud elemente ja RDF notatsiooni.

Andmesõnastiku termini kirjeldus

#	Elemendi nimetus	Määratlus ja kasutamine	Kohustuslik	Näide	OWL/RDF element[45]
1	Termin	Andmesõnastiku termini inimloetav kuju. Reeglina saadakse ärisõnastikust.	Jah	Ehitise suletud netopind	rdfs:label
2	Termini määratlus	Termini vabatekstiline seletus. Reeglina saadakse ärisõnastikust või luuakse andmekirjeldamise käigus.	Ei	Hoone suletud netopind on kõigi korruste suletud netopindade summa	dct:description
3	Termini URI	Termini unikaalne identifikaator andmesõnastikus (URI). Märkus: URI võetakse andmesõnastike jaoks kasutusele koos RIHAKEse juurutamisega.	Jah	http://ehr.ee/ehitised/dataDictionary/v2/ term:ehitiseNetopind	rdf:ID
4	Termini staatus	Märge selle kohta, kas termin on kinnitatud ärisõnastikku või veel mitte. Kasutatakse ühte staatus valikust: kinnitatud; taotletud; kinnitamata.	Ei	Kinnitatud	-
5	Termini kehtivus	Märge selle kohta, kas termin on veel kasutusel (Jah/Ei). Termin võib olla kasutusest võetud mitmesugustel põhjustel, nt. andmekoosseisu muutus.	Jah	Jah; Ei	-
6	Termini allikas	Viide termini kirjelduse allikale, juhul kui allikas ei ole ärisõnastik. Viidatakse näiteks teisele sõnastikule, standardile või kirjandusele. Kasutatakse vajadusel terminite ja nende määratluse allika näitamiseks, kui taotletakse andmekirjeldamise käigus loodud uute terminite lisamist ärisõnastikku.	Ei	GSIM versioon 1.2 https://statswiki.unece.org/display/clickablegsim/Data+Set	rdfs:isDefinedBy

[45] Andmesõnastiku terminite kirjeldamisel kasutatakse võimalusel W3C Web Ontology Language (OWL, https://www.w3.org/OWL/) standardis terminite (individual) ja omaduste (properties) kirjeldamiseks defineeritud elemente ja RDF notatsiooni.

7	Termini loomiskuupäev	Kuupäev, mil termini kirjeldus lisati andmesõnastikku (vormingus AAAA-KK-PP).	Jah	2020-02-16	dct:created
8	Termini muutmiskuupäev	Kuupäev, mil viimati termini kirjeldust muudeti (vormingus AAAA-KK-PP).	Ei	2020-12-11	dct:modified
9	Termini seosed	Viide terminile ärisõnastikus (URI). Märkus: URI võetakse andmesõnastike jaoks kasutusele koos RIHAKEse juurutamisega.	Ei	http://ehr.ee/ehitised/BuildingsOntology/v3/ehitiseNetopind	dct:relation
10	Märkused	Vabatekstiline kirjeldus termini kohta.	Ei	Ilmselt muutub haldusreformi tulemusena; Kasutusotstarve ebamäärane	-