Riigi Infosüsteemi Amet

AVAANDMETE LOOMISE JA AVALDAMISE JUHEND

Versioon 1.1

Tallinn 2015, uuendatud 2016

Sisukord

1. Dokumendi käsitlusala

Käesolev dokument kirjeldab millised andmed tuleb avaldada kui avaandmed ja millistest põhimõtetest lähtuda avaandmete avaldamisel.

Dokument ei kirjelda seda, milliseid konkreetseid vahendeid peab kasutama avaandmete loomisel ja avaldamisel, vaid seda, millistest põhimõtetest lähtudes luua avaandmeid, kuidas struktureerida ning versioneerida avaandmeid ja kuidas need teha kättesaadavaks kasutajatele.

Juhend on mõeldud kõigile andmete omanikele, kelle valduses on andmeid, mida ennast või mille teisendusi saab käsitleda kui avaandmeid ja avaandmete kasutajatele. Konkreetsemalt on juhend loodud abistamaks töötajaid, kes tegelevad avaandmete avaldamise planeerimise ja andmete avaldamise protsesside üles ehitamisega.

2. Alusmaterjalid

Käesoleva dokumendi kirjutamisel on lähtutud järgmistest alusmaterjalidest:

1. **Eesti Vabariigi Põhiseadus**, §44 - üldiseks kasutamiseks levitatava informatsiooni vabast kasutamisest (www.riigiteataja.ee/akt/115052015002)

2. **Avaandmete roheline raamat**, mis sisaldab Eesti Vabariigi avaandmete poliitika süstemaatilise käsitluse (https://opendata.riik.ee/sites/default/files/manuals/avaliku-teabe-masinloetava-avalikustamise-roheline-raamat-20141125_0.odt)

3. **Avaliku teabe seadus** (AvTS), mille eesmärgiks on tagada üldiseks kasutamiseks mõeldud teabele avalikkuse ja igaühe juurdepääs ning luua võimalused avalikkuse kontrolliks avalike ülesannete täitmise üle. (https://www.riigiteataja.ee/akt/112072014033)

4. **Isikuandmete kaitse seadus** (IKS), mille eesmärk on kaitsta isikuandmete töötlemisel füüsilise isiku põhiõigusi ja -vabadusi, eelkõige õigust eraelu puutumatusele (www.riigiteataja.ee/akt/112072014051)

5. Dokumendis lähtutakse ka Eesti Vabariigi **avaandmete portaali** (opendata.riik.ee) käesoleva dokumendi loomise ajal kehtivast struktuurist ja funktsionaalsusest ning projekti SharePSI 2.0 (www.w3.org/2013/share-psi/) käigus kogutud parimatest praktikatest.

3. Avaandmete olemus

Avaandmed on andmed, millele igaüks omab piiranguteta õigusi andmete lugemiseks, töötlemiseks sh. modifitseerimiseks ja taas esitamiseks. Kitsamas mõttes mõistetakse käesolevas dokumendis avaandmete all avalikke, riigi ja kohaliku omavalitsuse tasandi valitsemise tulemusena tekkivaid mitte-konfidentsiaalseid andmeid. Laiemas

mõttes käsitletakse siin dokumendis avaandmeid kui mistahes avalikuks, piiranguteta kasutamiseks mõeldud ja vabalt kätte saadavaid andmeid, sõltumata nende tekke allikast, mis on avalikuks tehtud nende andmete omaniku poolt.

Avaandmete eesmärk on baasi loomine erinevaid valdkondi ühendavate arvutiseeritud rakenduste loomiseks ja seeläbi ühiskonna kui terviku koostalitlusvõime parendamine ning riigivalitsemise läbipaistvuse tõstmine.

Avaandmete planeerimisel, projekteerimisel, loomisel ja avaldamisel tuleb lähtuda järgmistest kriteeriumitest:

a) eraldatus – **üldjuhul** tuleb avaandmed luua selliselt, et nad on lahus tööandmetest st. avaandmete avaldamine ei peaks üldjuhul olema lahendatud viisil, kus avaandmete avaldamiseks piiratakse tööandmetele juurdepääsu avaliku kasutaja jaoks lihtsalt andmete lugemisrežiimi (*read only*) piiranguga. Viimati kirjeldatud viisi (avaandmete käsitlemist kui tööandmete lugemisrežiimis avaldamist) kasutamisel tuleb hoolikalt kaalutleda kõiki riske, mis võivad kahjustada sellisel viisil avaandmeid jagava infosüsteemi käideldavust – riskid tarkvara uuendamisel tekkivatest võimalikest (avastamata) turvaaukudest, avalikust võrgust lähtuv ettearvamatu serveri koormuse tõus, mis häirib infosüsteemi igapäevast tööalast kasutamist jms.

b) kättesaadavus – andmed peavad olema võrdselt kättesaadavad kõigile sõltumata sellest, milline on andmete kasutaja residentsus võrreldes andmete residentsusega, milline on andmete kasutaja juriidiline staatus või milline on andmete kasutamise eesmärk.

c) juurdepääs – andmed peavad olema juurdepääsetavad viisil, mis teeb need lihtsalt kättesaadavaks võimalikult laiale kasutajaskonnale. Üldjuhul mõeldakse selle all andmete kätte saadavust interneti kaudu.

d) perioodilisus – samasisulisi andmeid tuleb esitada omavahel võrreldavate perioodide kaupa; kui samade andmete perioodilisust aja jooksul muudetakse tuleb seda võimalusel teha selliselt, et võrreldavus varem avaldatud andmetega ei kaoks.

e) järjepidevus – sama koosseisuga avaandmeid tuleb avaldada kogu ajatelge katvana (ilma aukudeta).

f) terviklikkus – andmed peavad olema struktuurilt ja sisult terviklikud st. andmed peavad olema struktureeritud selliselt, et kokku kuuluvad andmed on omavahel ilmutatult seotud ja andmetes ei ole viiteid sellistele andmetele, mis ei ole andmete kasutajale kätte saadavad avaandmetena.

g) terviklus – avaandmete füüsiliselt erinevad kogumid, mis kõigi tunnuste järgi on tuvastatavad samade andmete kogumina, ei tohi olla variantsed ja ajas muutuvad.

h) esitusviis – avaandmed esitatakse mingi üldtunnustatud, vaba litsentsiga kaetud formaadiga formaliseeritud kujul; andmete esitamisel tuleb taotleda formaadi lihtsust ja esitatavate andmete mugavat töödeldavust.

i) mõistetavus – andmed peavad olema kirjeldatud sellisel tasemel, et vajalike teadmiste ja oskuste olemasolul oleksid need kõigi nende andmete kasutajate jaoks semantiliselt ühte moodi mõistetavad; sisuliselt tähendab see seda, et avaandmetena on kätte saadavad ka kõikide andmete (andmeelementide) semantilised kirjeldused.

j) litsentseerimine – avaandmed tuleb üldjuhul esitada selliste litsentsitingimuste all, mis võimaldavad nende andmete töötlemist sh. nende andmete rikastamist teistest mistahes andmekogumitest pärit andmetega, teiste mistahes andmekogumite rikastamist nende andmete alusel ja nii lähteandmete kui tuletatud andmete nii tasuta kui tasulist esitamist.

k) proaktiivsus vs. tellimused – alustavalt peab andmete valdaja andmeid avaldama oma initsiatiivist lähtuvalt ja enda poolt kehtestatud korra alusel ja seaduses kehtestatud mahus. Proaktiivsuse nõue tuleneb Avaliku teabe seaduse §28-st (AvTs §28-s on esitatud loend avalikustatavatest andmetest), mis kohustab teabe valdajaid avalikustama nende käsutuses olevaid andmeid. Tellimuse tekkimisel konkreetse struktuuri, sisulise hõlmavuse, ajalise katvuse ja avaldamise sagedusega andmetele lepitakse andmete avaldamise detailides kokku andmete kasutajatega.

l) anonümiseerimine – kui seaduses ei ole öeldud teisiti, st seadusega on määratud isikustatud andmete avaldamise kohustus, ei tohi avaandmetena esitatavad andmed olla seostatud/seostatavad isikuga st. üldjuhul tuleb avaandmed luua viisil, kus puuduvad viited konkreetsetele isikutele või temaga seotud asjadele, toimingutele või asukohtadele nii, et nende andmete kõrvutamisel ja võrdlemisel mistahes teiste andmekogumitega (ka avaandmetega) ei oleks võimalik avaandmetes esitatud andmeid seostada konkreetsete isikutega. See kõik puudutab inimese kui eraisikuga seotud andmeid ja ei puuduta inimese kui ametnikuga seotud andmeid. Seda muidugi juhul, kui seadus ei ütle konkreetsete andmete kohta teisiti.

m) konfidentsiaalsus – avaandmed ei tohi sisalda konfidentsiaalseid andmeid; konfidentsiaalseteks loetakse andmed, mis on kaetud riigisaladuse kaitsega, mis on kuulutatud ametialaseks kasutatamiseks (AK) ettenähtud andmeteks või mis on kaetud seadustest tulenevate juurdepääsu piirangutega.

n) andmete õigsus – andmete sisu ja õigsuse eest vastutab nende esimene sellises komplekssuses esitaja so. andmete omanik.

4. Avaandmed

Avaandmete sisu kirjeldub kui kõikide avalike teenuste osutamisega seotud andmed, mis on anonümiseeritud tasemeni, kus neid ei ole võimalik seostada üksikisikutega ja mis ei ole kaetud riigisaladuse kaitsega või ametialase kasutuspiiranguga (AK) ja millele ei rakendu seadustest tulenevaid muid juurdepääsupiiranguid. Samas, selle üldistatud definitsiooni taha, on peidetud lai võrk üsna olulisi probleeme, mille terviklik mõistmine ja selles lähtuv meetodite valik tagab korrektselt üles ehitatud ja toimivad avaandmed.

Käesolevas jaotises vaadeldakse avaandmete struktureerimise, loomise, avaldamise ja haldamise erinevaid aspekte.

4.1. Avaandmete piiritlemine

Rääkides avaandmetest tuleb mõista, et avalike teenuste osutamise käigus tekkivad tööandmed ei ole **enamikel juhtudel** automaatselt avaandmed. Seda peamiselt kolmel põhjusel:

1. tööandmed ei ole anonümiseeritud
2. tööandmed võivad sisaldada riigisaladuse, ametialase kasutuspiirangutega (AK) kaitstud ja seadustest tulenevate juurdepääsupiirangutega andmeid
3. tööandmed ei saa olla kaetud avaldamist ja igaühe poolt töötlemist ning taas esitamist võimaldavate litsentsi tingimustega

See kõik muidugi ei tähenda seda, et tööandmed ei võiks olla oma sisult samastatavad avaandmetega – kõigis tööandmetes ei pruugi üldsegi sisalduda punktides 1 ja 2 kirjeldatud piirangutega andmeid.

Eesti Vabariigi tingimustes ei ole seotus litsentsiga eriliseks piiranguks. Seda selle pärast, et Eesti Vabariigi seaduste järgi on kõik avalike teenuste osutamise protsessi käigus tekkivad andmed, mille kohta ei ole seadusega seatud avaldamise piiranguid, vabalt levitatavad ja piirangutega kasutatavad andmed, mida ei pea kuidagi litsentseerima vaid otse vastupidi – seadus kohustab neid avaldama (AvTS §28).

Ameerika Ühendriikidest ja eriti Suurbritanniast levinud andmete litsentseerimis- ja kasutustava kohasel peab andmete kasutaja ilma litsentsita andmete kasutamiseks küsima andmete valdajalt luba isegi juhul, kui need andmed on avalikult kätte saadavad. Eesti Vabariigi tingimustes on see luba seadustega antud ettehaaravalt. Seega otsest kohustus või vajadust andmete mingi üldlevinud litsentsiga sidumiseks meil ei ole. Samas, vältimaks teistest riikidest pärit kasutajate segadusse ajamist, võiks andmete avaldamise kohtades viidata näiteks CC0 (Creative Commons Zero; https://creativecommons.org/about/cc0) litsentsile või mõnele teisele sarnasele litsentsile.

Mõnedel juhtudel tuleb siiski kaaluda rangemate litsentsitingimuste kasutamist. Näiteks seadusetekstide (ja samaväärsete andmete) avaldamisel avaandmetena võiks kasutada litsentsi, mis lubab küll andmete kasutamist taasavaldamisel ja teiste andmete rikastamisel, kuid piirab antud andmete esitamist muudetud kujul. Common Creative perekonnast pärit litsentsidest sobib selleks CC ND litsents (Creative Commons No Derivates).

4.2. Avaandmete avaldamise viis

Lähtudes eelmistes jaotistes kirjutatust on avaandmeid võimalik avaldada põhimõtteliselt kahel erineval viisil:

1. tööandmetest eraldi seisvate andmekogumitena, mis on spetsiaalselt avaandmetena ettevalmistatud
2. *read only* režiimis juurdepääsuna tööandmetele

Teist varianti ei saa kasutada kindlasti sellistel juhtudel, kus tööandmed sisaldavad isikuandmeid, riigisaladust, ametialaseks kasutamiseks (AK) määratud andmeid või seadustest tulenevate juurdepääsupiirangutega kaitstud andmeid. Sellisel juhul tuleb tööandmed „puhastada“ avaldamispiiranguga andmetest ja esitada eraldi seisva andmekogumina. Isegi juhul, kui päringute filtritega õnnestuks välistada avaldamispiiranguga andmete sattumine avaldatavate andmete hulka, on selle meetodi kasutamine ohtlik, kuna mingi programmeerimisvea või liidese turvaaugu tõttu võivad andmed, mida ei tohi avaldada, saada avalikuks. Loomulikult jääb selle variandi kasutamise otsustus andmete avaldaja teha – kui suurt riski ta on nõus võtma.

Teise variandi kasutamine on õigustatum siis, kui tööandmed uuenevad väga tihti ja seega tuleks tihti moodustada ka avaldatavate andmete kogumeid. Kui sellega kaasneb veel ka avaldamispiirangutega andmete puudumine siis võibki otsustada just teise variandi kasuks.

Seega esimese variandi kasuks on mõttekas otsustada siis, kui tööandmetes on avaldamispiiranguga andmeid ja teise kasuks siis, kui tööandmetes avaldamispiiranguga andmeid ei ole ja andmete muutumise sagedus on suur.

Teise variandi kasutamisel ei tohi ära unustada seda, et me avame oma infosüsteemi välisele kasutajale, ja sellest tulenevad olulised riskid meie infosüsteemile:

1. baastarkvara uuendamise tulemusel võivad tekkida infosüsteemi turvaaugud, mille kaudu on võimalik rünnata nii andmete omaniku käsutuses olevaid andmeid kui ka tehnoloogiaid. Kuna tegemist on tööandmete ründega siis võib tekkida oluline andmekadu ja seda just kiiresti uuenevate andmete puhul. Lisaks sellele paneb see igakordsel tarkvarauuendusel lisakoormuse juurdepääsu testidele.
2. funktsionaalse tarkavara uuendamisel ja sellest tulenevatest andmebaasimuudatustest tingituna võib toimuda andmete loogika selline muutus, mis võib muuta avaldamispiiranguga andmed nähtavaks selleks mitte volitatud isikutele läbi avaandmete liidese. Seda eriti just dünaamiliselt üles ehitatud liideste korral. Seega suureneb tarkvara uuendamisel jäälegi koormus juurdepääsu testide tegemisele.
3. Avalikust võrgust tuleva koormuse tõusu tõttu võib olla häiritud infosüsteemi kasutajate igapäevane töö

Avaandmete avaldamisel on soovitav kasutada esimest varianti, kuna sellega kaasneb vähem riske – avalikust võrgust tulevaid kasutajaid ei lasta ligi tööandmetele ja töösüsteemidele ning avaldatavate andmete struktuur ja sisu on range kontrolli all.

4.3. Andmete anonümiseerimine vs. sünonümiseerimine

Kui seadus ei ütle konkreetsete andmete kohta teisiti, tuleb kõik avaandmetena esitatavad andmed anonümiseerida sellisel tasemel, et erinevate andmete kõrvutamisel ja ühendamisel ei oleks võimalik taastada seost konkreetse isikuga, kellega andmed tööbaasis on seotud. Seejuures tuleb andmed just anonümiseerida mitte sünonümiseerida. Need kaks tegevust (anonümiseerimine ja sünonümiseerimine) on oma olemuselt täiesti erinevad protsessid.

Sünonümiseerimine tähendab seda, et sama isiku identifitseerivad andmed asendatakse kõigis avaandmete andmekogumites sama koodiga ja selle koodi alusel on võimalik kõik erinevate andmekogumite andmed omavahel kokku siduda ilma konkreetset isikut teadmata. Samas, kui piisavalt palju andmeid on omavahel kokku seotavad, siis on võimalik üsna suure tõenäosusega ka isiku identiteedi taastamine. Selleks tuleb lihtsalt enda käsutuses olevate isikuandmete alusel analüüsida kokku kogutud ja ühendatud avaandmeid ja oma isikute kogumi piires (näiteks klientide register) on võimalik tuvastada kellele millised andmed avaandmetes kuuluvad. Seepärast on sünonümiseerimine avaandmete esitamisel keelatud.

Anonümiseerimine tähendab tööandmetest avaandmete loomise protsessis isiku identiteedi kustutamist kõikidest loodavatest avaandmete andmekogumitest. See

tähendab seda, et erinevate avaandmete andmekogumite andmed muutuvad oma vahel seostamatuteks ja isiku tuvastamine andmete analüüsiga muutub võimatuks.

See piirang on väga oluline ja seda tuleb järgida pretsedenditult. Põhjus on selles, et ühe andmeavaldaja viga selles vallas võib tekitada isiku andmete tuvastamatuse reegli rikkumise teiste andmeavaldajate juures. Selle reegli rikkumine mitme andmeavaldaja juures võib tekitada kuhjumise ja tingida olukorra, kus üle kõigi andeavaldajate avaandmete võib tekkida võimalus isiku tuvastamatuse reegli rikkumiseks.

4.4. Anonümiseerimine ja identifitseerivad andmed

Andmete anonümiseerimist tuleb rakendada igal pool, kus Isikuandmete kaitse seaduse järgi tekkib võimalus isiku eraelu kaitse riiveks. Seda reeglit on võimalik mitte rakendada vaid juhul, kui kellegi õigustatud huvi kaalub isiku eraelu riive üles.

Isikuga ametialaselt seotud andmed ei kuulu üldjuhul anonümiseerimisele, kui seadus ei sätesta teisiti.

Isikut identifitseerivaid andmeid on oluliselt rohkem, kui seda on tema nimed ja isikukood. Sellisteks andmeteks on isiku mistahes dokumentide numbrid (isikutunnistus, pass, juhiluba, pensioni tunnistus, sünnitunnistus jne.), isiku kohta koostatud dokumentide numbrid (avaldused, korraldused, aktid, load, otsused jne.), isiku aadressid, isikuga seotud asutuste/ettevõtete nimed, autode numbrid, autode kerenumbrid jpm., mis erinevate andmete omanike lõikes on väga erinevad. Seepärast tuleb anonümiseerimine protsessi käigus “kaotada” ka need andmed.

“Kaotamine” ei tähenda siin kõigil juhtudel nende andmete kustutamist. Vahetute identifikaatorite puhul, mis on olemuselt sarnased isikukoodiga, nagu näiteks dokumentide numbrid ja isiku kohta koostatud dokumentide numbrid, autode numbrid, on muidugi andmete kustutamine vältimatu. Samas aadressiandmete puhul ei ole see kohustuslik vaid kustutamise võib asendada üldistamisega. See tähendab seda, et konkreetne aadress asendatakse näiteks viitega linnaosale, asulale, vallale või maakonnale.

Kõik **andmete üldistamised** tuleb läbi viia kindlate reeglite alusel, nii, et tagatud oleks peamine eesmärk – üldistatud andmete alusel ei tohi olla võimalik taastada isiku identiteeti. Näiteks kui me oleme kehtestanud aadressi üldistamise reegliks linnas linnaosa täpsuse, asulas asula täpsuse, külas küla täpsuse ja maal valla täpsuse, siis kui kusagil väikeses külas elab ainult üks inimene, siis ei “peida” küla täpsusega üldistus tema identiteeti. Asi pole parem ka siis, kui selliseid inimesi on kolm – teada olevate isikute ring on piisavalt väike, et analüüsi tulemusena tuvastada tegelikud isikud. Sisuliselt tähendab see seda, et pärast avaandmete andmekogumi loomist tuleb teostada enne selle avaldamist reeglipõhine andmete kontroll, mis peab sellised situatsioonid tuvastama ja välistama. Kontrollida tuleb seda, kas antud andmekogumis on mõni väärtus, mis on määratud väga väikesele kogumile andmetele ja kui selline situatsioon leitakse, siis suurendada selle omaduse üldistuse taset. Meie näite puhul siis üldistuse hajutamine valla tasemelt maakonna tasemele.

Identifitseerivate andmetena võivad toimida ka sellised andmed, mis esmapilgul sellena üldse ei klassifitseeru. Näiteks massiliselt tarbitavate ravimite nimetused seda kindlasti pole. Samas on väga unikaalseid ravimeid, mida tarbivad ainult üksikud inimesed. Unikaalsete ravimite tarbimist tingivad jällegi unikaalsed haigused, mis on jällegi isikut üsna täpselt määravaks asjaks. Samuti võib selliseks identifitseerivateks andmeelemendiks olla näiteks väga unikaalne auto mark. Sellistel juhtudel hakkab

siin tööle sama loogika, mida selgitati elukoha üldistamise näite korral – harva määratav ravim või harva esinev auto mark määravad piisavalt väikese ringi inimesi, kelle isiku tuvastamine järelanalüüsiga on piisavalt lihtne.

Sellistel juhtudel tuleb kasutada **andmete asendamist** – vähe kasutatavad andmed tuleb asendada selliselt, et samade väärtustega andmeid tekiks rohkem. (Näiteks: auto mark "Hummer H2 Luxury" asendatakse fraasiga "luksusauto"; ravim "Riociguat" asendatakse fraasiga "eriravim"). Kui see ei õnnestu (st. unikaalsus säilib ka pärast asendamisi) siis tuleb terve andmekomplekt valimist kustutada. Viimane reegel kehtib kõikidel juhtudel, kus ühe andmekirje piires ei õnnestu andmeid piisavalt anonümiseerida.

Identiteedi taastamist lihtsustavad andmed võivad olla peidetud ka andmeelementide **kombinatsioonide** taha. Näiteks kui sama andmekomplekti raames on valla tasemele üldistatud aadress ja "luksusautoks" asendatud "Hummer H2 Luxury" ning selles vallas on ainult üks "luksusauto", siis pole raske ära arvata, milline isik selle taga on. Sellisel juhul on kaks võimalust – kas üldistada aadressi veel kõrgemale tasemele või siis kustutada see andmekomplekt valimist üldse. Otsustavaks kahe variandi vahel saab see, kui palju sama moodi mõistetavaid ridu tehtud muudatuste tulemusel andmekogumisse tekkib.

Samasuguse näite saaks konstrueerida juhul kui külas on üks elanik ja on teada, et selle küla elanik ostis mingit haruldast ravimit või on tal mingi haruldane haigus. Siin on seos samuti ühene ja lahendada tuleb seda analoogiliselt eelmise näitega.

Vajalikud tegevused:

1. Vahetult identifitseerivate andmete (nt. isikukood, nimed, dokumendinumbrid jne.) määratlemine ja nende avaandmetesse lisamise keelamine
2. Üldistatavate andmete (nt.: aadress jms.) määratlemine ja üldistamise reeglite kehtestamine
3. Harva esinevate, asendatavate andmete (nt.: unikaalsed ravimi nimed, harva esinevate haiguste nimed, harva esinevad automargid jne.) määratlemine ja asendusreeglite kirjeldamine
4. Isiku tuvastamatuse reegli rikkumist võimaldavate andmekoosluste (nt.: asukoht ja auto mark jms.) määratlemine
5. Kontrollireeglite kirjeldamine üldistatud andmete (p. 2), asendatud andmete (p. 3) ja isiku tuvastamatuse reegli rikkumist võimaldavate andmekoosluste (p. 4) kohta ja isiku tuvastamatuse reegli rikkumise kõrvaldamise reeglite kirjeldamine.
6. Eelpool kirjeldatud reeglite (p. 1, p. 2, p. 3, p. 4 ja p. 5) täitmise tarkvaraline tagamine.

4.5. Andmete koosseis ja variantsus

Avaandmete ühe andmekogumi koosseis (struktuur) ei tohi olla väga lai st. avaandmete andmekogumite loomisel ei tohiks olla rakendatud põhimõte, kus "kõik" sama valdkonna andmed on avaldatud ühes andmekogumis. Selleks on mitmeid põhjuseid:

1. laiaulatuslike andmestruktuuride puhul on andmete maht suur ja alla laetavate andmete maht suureneb ilmaasjata, sest erinevate ülesannete lahendamiseks vajatakse erinevaid andmekogumeid, aga mitte kunagi kõiki andmeid
2. andmekogumite struktuuri piiramisega on võimalik luua sihtorienteeritud andmekogumeid, mille semantilise määratluse (sisu) kaudu on võimalik suunata kogukondi (anda ideid) riigi ja omavalitsuse seisukohalt vajalike rakenduste loomisele.
3. võimaldab hoida juba loodud andmekogumi struktuuri läbi kõigi versioonide (läbi aja) stabiilsena – uute andmete lisandumisel luuakse lihtsalt uued andmekogumid, millel on uus struktuur ja mis sisaldavad lisaks „vanadele andmetele“ ka „uusi andmeid“. Andmekogumite andmestruktuuride stabiilsuse hoidmine on äärmiselt oluline, kuna sõltuvalt andmekogumi atraktiivsusest võib sellega olla seotud mitmeid rakendusi ja andmekogumi struktuuri iga muudatus mõjutab nende toimimist.
4. Konkreetse huvi tekkimisel andmete vastu on võimalik luua mingi konkreetse kasutajate grupi jaoks sihtotstarbeliselt kasutatavaid andmekogumeid.

Selliste printsiibi rakendamine võib tunduda andmete valdaja „õiguste ahistamisena“ käsitleda oma andmeid „nii nagu ta tahab“, kuid kogu avaandmete avaldamise pikemaks eesmärgiks on kogukondi abistavate tarkvaraliste rakenduste loomine nende teenuste arendajate ja pakkujate poolt. Seega tuleb tagada keskkond, kus need rakendused saavad stabiilselt eksisteerida.

Vajalikud tegevused:

1. Planeerida tegevussuunad, mille arendamist oodatakse
2. Planeerida avaandmete andmekogumite semantikad ja nendele vastavad andestruktuurid selliselt, et kõik andmete valdaja käes olevad avalikud andmed oleks avaandmete erinevate andmekogumitega kaetud.
3. Vaadata olemas olevate andmekogumite struktuurid aeg-ajalt üle (ette määratud sagedusega) ja analüüsida uute struktuuridega andmekogumite loomise vajadust.
4. Suhelda andmete vajajatega ja arendada neile vajalike struktuuridega andmekogumeid.

4.6. Andmekogumite versioonid, ajaline määratlus, ajaline järjepidevus ja avaldamise perioodilisus

Oluline! Versioneerida saab ainult selliseid avaandmeid, mis on moodustatud tööandmetest eraldi seisvate andmekogumitena. Avaandmed, mis on lahendatud *read only* režiimis juurdepääsuna tööandmetele uuenevad samas kohas kus nad tekkivad, töö rütmilisusest tuleneva sagedusega. Seega, kogu käesolevas jaotises kirjeldatu puudutab ainult selliseid avaandmete kogumeid, mis on loodud tööandmetest eraldi seisvate andmekogumitena.

Kõik avaandmete andmekogumid peavad olema seotud selgelt määratletud ajavahemikuga. See tähendab seda, et andmete kasutaja peab teadma millise perioodi kohta käivad andmed on konkreetses andmekogumis.

Sama struktuuriga andmekogumit tuleb aja möödudes avaldada korduvalt st. samast andmekogumist tuleb teha aja möödudes uusi versioone. Sama andmekogumi iga uus versioon samastub eelmiste versioonidega struktuurilt. See tähendab seda, et andmekogumi iga uue versiooni struktuur on sama mis andmekogumi eelmistelgi versioonidel. Sama andmekogumi iga uus versioon eristub eelmistest versioonidest ainult ajavahemiku poolest, millesse kuuluvaid andmeid see sisaldab. See tähendab seda, et sama andmekogumi erinevate versioonide ajavahemikud ei tohi kattuda.

Sama andmekogumi erinevad versioonid peavad olema pikema perioodi jooksul loodud sama pikkusega ajavahemike kohta. See muudab need omavahel võrreldavateks ja hõlbustab oluliselt andmete paigutamist ajateljele. See tähendab seda, et sama andmekogumi uute versioonide väljastamiseks peab olema määratud sagedus so. ajavahemik, mille tagant uusi versioone avaldatakse.

Andmekogumi igasse versiooni kuuluvad andmed, mille ajaline periood algab eelmise versiooni ajaperioodi lõpu hetkele järgnevast hetkest. Perioodi lõpp määratakse vastavalt antud andmekogumi jaoks kirjeldatud reeglistikule.

Andmed, mis ei ole ajaliselt periodiseeritavad esitatakse hetke väärtustena. Ka sellisel juhul tuleb järgida andmete perioodilise väljaandmise printsiipi, kus andmeid avaldatakse pikema aja jooksul sama pika ajavahemiku tagant.

Oluline! Iga andmekogum (andmekogumi versioon) peaks sisaldama ilmutatud kujul andmeid selle perioodi kohta, millesse kuuluvaid andmeid see andmekogum sisaldab – vaadeldava perioodi algus ja lõpu aega.

Vajalikud tegevused:

1. Selgitada välja (uuring ja analüüs) erinevate andmete avaldamise mõistlik sagedus
2. Määrata kõikidele avaldatavatele andmekogumitele ajavahemikud, mille tagant antud andmekogumi uus versioon regulaarselt avaldatakse
3. Tagada kogu ajatelje katvus sama struktuuriga andmetega.

4.7. Esitlusvorming

Avaandmete esitamisel eraldiseisvate andmekogumitena valib andmete esitamise vormingud andmete esitaja. Üdiseks printsiibiks andmete esitusviisi valimisel on nende visuaalne arusaadavus ja töödeldavuse lihtsus ja kasutatava vormingu tuntus st. ei ole mõistlik kasutada vähelevinud vorminguid, kuna kasutajatel puudub nende kasutamise kogemus ja vahendid.

Soovitatavad esitlusvormingud on:

1. JSON - JavaScript Object Notation (http://json.org)
2. JSON-LD – JSON laiendus linkandmetega (http://json-ld.org)
3. XML – Extensible Markup Language (http://www.w3.org/TR/xml/)
4. CSV – Comma separated values (http://data.okfn.org/doc/csv)

Võimalusel on soovitav kasutada kõigis avaldatavates andmekogumites sama vormingut. Statistilisi andmeid on otstarbekas avaldada linkandmetena. Sellisel viisil saab andmekogumeid ühendada ühiselt kasutatavate kontseptsioonide abil (näiteks kasutades RDF Data Cube sõnastikku).

Avaandmed võivad olla esitatud ka veebiteenuse või mõne muu interaktiivse liidese kaudu (nt. andmebaasi liidese kaudu). Sellisel juhul määrab andmete esitlusvormingu konkreetne liides ja rakendatakse kasutatavale liidesele iseloomulikke andmete esitlusvorminguid.

Vajalikud tegevused:

1. Valida välja kasutatavad vormingud
2. Määrata kõigile andmekogumitele esitusvorming.

4.8. Andmeelementide vormindus

Avaandmetes kasutatavad andmete vormindused peaavad tagama andmete ühese andmetüübi interpreteeritavuse. Erinevate andmetüüpide puhul tuleb rakendada erineva tugevusega vormindamist.

Andmetüüp	**Vormindus**
Tekst	vormindamata
Arv	Lubatud sümbolid: 0...9; , (koma); – (miinus) Täis- ja kümnendosa eraldamiseks kasutatakse koma. Järkude eraldamist erisümbolitega (tuhandelised) ei kasutata. Vahetult negatiivsete arvude ees on „miinus“ (ilma arvu ette jääva eraldussümbolita) Täisarvudel kümnendosa ja seega ka koma puudub. (N: 1,23; 2)
Kuupäev	Kuupäev esitatakse kujul „AAAA-KK-PP“ ilma eraldajateta, kus: „AAAA“ on neljakohaline aastaarv, „mm“ on kahekohaline (vajadusel eesnulliga) kuu number ja „PP“ on kahekohaline (vajadusel eesnulliga) päeva number (N: 2015-08-04)
Kellaaeg	Kellaaeg esitatakse kujul „tt:mm:ss[,nnn]“ ilma eraldajateta, kus: „tt“ on kahe kohaline (vajadusel eesnulliga) tunni number vahemikus 00-24, „mm“ on kahe kohaline (vajadusel eesnulliga) minutite arv, mis on möödunud täistunnist, vahemikus 00-59, „ss“ on kahe kohaline (vajadusel eesnulliga) sekundite arv, mis on möödunud täisminutist, vahemikus 00-60. Sekundi number 60 on võimalik, kui toimub kellaaja korrigeerimine 1 sekundi (nn. *leap second*) võrra. „[,nnn]“ on kuni kolme kohaline (mitte kohustuslik)

	millisekundite arv, mis on möödunud täissekundist. (N: 18:27:04; 18:27:04,372)
Andmetüüp	**Vormindus**
Kuupäev ja kellaaeg	Kuupäeva ja kellaaja formaat, mis on sidurdatud üle sümboli „T“ (N: 2015-08-04T15:04:23)
Kestvus	Kestvus esitatakse kujul „P[[<n>Y][<n>M][<n>D]][T[<n>H][<n>M][<n>S]]“ või P<n>W, kus „[…]“ tähistab kombinatsiooni puudumise võimalikkust, „P“ - tähistab kestvuse määrangu algust, „<n>“ tähendab kestvuse pikkust ”Y“, „M“, „D“, „H“, „M“ (pärast „T“-d, mis tähistab kellaajavormingu algust), „S“ ja „W“ tähistavad vastavalt kestust aastates (Y), kuudes (M), päevades (D), tundides (H), minutites (M pärast T-d, mis tähistab kellaajavormingu algust), sekundites (S) ja nädalates (W)
Ajaintervall	Ajaintervalli kirjeldamiseks kasutatakse kahte sama täpsusega (kuupäev, kellaaeg või kuupäev ja kellaaeg) ajavormingut (algus ja lõpp), mis eraldatakse sümboliga „ / “ Ajaintervalli võib esitada ka kahe eraldi seisva andmeväljana – algus ja lõpp. Ka sellisel juhul peaksid need olema sama täpsusega
Klassifikaatoriväärtused	Kui andmeelement on seotud riikliku klassifikaatoriga, siis peab iga kord olema esitatud kolmik: klassifikaatori kood, klassifikaatori rea kood, rea nimetus.

Vajalikud tegevused:

1. Jaotises kirjeldatud formaatide kasutamine

4.9. Andmete asukoht ja leitavus

Konkreetne andmete füüsiline asukoht, mille kaudu avaandmeid avaldatakse, ei ole üldse oluline. Olulised on viis omadust, mis andmete avaldamise asukohta iseloomustavad:

1. andmete avaldamise asukoht (URL) peab kõigile soovijatele olema lihtsalt leitav,

2. andmete asukoht (URL) peab olema pikema perioodi jooksul sama või peab samas asukohas (URL) asuma viit, mis osutab andmete tegelikule asukohale (URL)
3. andmete asukoht peab olema varustatud nii inim- kui masinloetava sisukorraga (need võivad ka kokku langeda),
4. juhul kui avaanded on avaldatud lahus tööandmetest (st. eraldi seisva andmekogumina) siis andmete avaldamise asukohas andmeid muuta ei tohi – säilima peab juba avaldatud andmete adekvaatsus; andmete parandused tuleb esitada eraldi, ära tuntavate andmekogumitena.
5. andmed peavad olema kätte saadavad mõistliku pidevusega; andmete asukoha töövõime ja kasutatavus (võime esitada andmeid) peavad olema tagatud sellise pidevusega, mis tagab nende andmete eesmärgipärase kasutuse.

Lihtne leitavus tähendab seda, et avaandmete andmekogud asuvad alla laadimiseks mingil avalikul internetiaadressil või on need kätte saadavad sellel aadressil asuva rakenduse kaudu.

Ilmselt ei ole pikemat ajahorisonti vaadates võimalik andmeid lõputult hoida samas kohas. See tähendab seda, et olude muutudes võib tekkida vajadus paigutada andmed ühest füüsilisest keskkonnast teise ja seega võib muutuda ka nende asukoha aadress. Seepärast on ilmselt otstarbekas hoida muutmatus kohas viita andmetele, mille asukoht võib muutuda. Sellise viida võib paigutada mitmesse kohta.

Selleks, et võimaldada ka nende viitade asukoha muutmine on mõistlik iga viida juures hoida viitasid ka teiste sarnaste viitade asukohale. Sellise käitumise puhul saab pikema ajaperioodi jooksul muuta, ilma andmete kasutajate jaoks ligipääsu aadresse ära kaotamata, ka kõikide juurdepääsupunktide asukohad. Seda tuleb teha ükshaaval ja kirjeldades „paigale jäävate“ juurdepääsupunktide juures ära uute ligipääsupunktide viidad ja eemaldades sealt vanad, mitte kehtivad viidad.

Kui avaldatavad andmed uuenevad korra ööpäevas siis on lubatud andmete avaldamise pidevuses pikemad katkestused kui sellisel juhul mil avaldatavad andmed uuenevad korra tunnis. Esimesel juhul peab andmete tarbija olema suuteline uusi andmeid lugema ühe ööpäeva jooksul. Teisel juhul peab andmete tarbija olema suuteline lugema andmeid korra tunnis. Vastasel korral (kui andmete lugemise selline sagedus ei ole tagatud) muutub küsitavaks andmete tarbija poolt pakutava teenuse toimimine. Seega, mida tihemini uuenevad avaldatavad andmed, seda kõrgemal tasemel peab olema ka avaandmete töövõime ja kasutatavus.

Andmete asukoha kasutajaõigused peavad olema loodud viisil, mis välistavad seal talletatavate andmekogumite muutmise. See välistab andmete pahatahtliku rikkumise. Andmekogumeid sealt võib ainult kustutada ja lisada sinna uusi andmekogumeid. Seda kõike vaid haldaja õigustes.

Tagatud peab olema pidev andmete säilivus (olemasolu) ja juurdepääs (kättesaadavus). See on kasutajate poolt andmete usaldamise nurgakivi ja seega üheks põhialuseks nendel andmetel baseeruvate rakenduste tekkimiseks. Kui kasutajad näevad, et andmeid ei ole võimalik kindlalt kätte saada, ei ehitata sinna peale ka rakendusi. See tähendab seda, et keskkond peab olema ehitatud piisavalt tõrkekindlana ja tagatud peab olema andmete taastamine. Kuna tegemist on avaandmetega (piiranguteta kasutatavate andmetega) võib andmete avaldamise keskkonnana kasutada mõnda eksisteerivat, võimalikult odavat või tasuta pilveteenust.

Vajalikud tegevused:

1. Andmete avaldamise asukohtade valimine ja administreerivate kasutajate kirjeldamine
2. Avalike viitade kirjeldamine andmete sisukorra asukohale
3. Andmete avaldamiskorra loomine ja avaldamine
4. Uute andmete üles panemine ja aegunud andmete kustutamine/muutmine

4.10. Andmete samastatavus

Avaandmed peavad olema jagatud erinevateks andmekogumiteks ja moodustama üksteisest lahutatud seeriaid st. versioonide jadasid. Sama andmekogumi kõik versioonid on alati kõik sama struktuuriga.

Samas võib eksisteerida ka paralleelseid andmekogumite seeriaid st. on olemas andmekogumid, mis sisaldavad põhimõtteliselt samu andmeid, kuid mille andmekooslus on üksteisest natuke erinev. Näiteks on juba eksisteerivale andmekogumi seeriale kõrvale tehtud uus, mis sisaldab põhimõtteliselt andmeid samade asjade kohta ja need on ligilähedaselt sama struktuuriga, kuid sinna on kas listaud uusi andmeelemente või siis eemaldatud neid sealt (vt p. 4.5).

Arvestades seda, et paljudel juhtudel toimub andmete töötlus programselt, on mõistlik luua andmekogumite nimede süsteem, mis võimaldaks määrata sama andmekogumi kõik versioonid (versioonide järgnevuse), sama andmekogumiga samatähenduslikud, kuid erineva andmestruktuuriga andmekogumid ja eristada neid täiesti erineva struktuuriga andmekogumitest. Selliseks eristamiseks sobib väga hästi andmekogumi füüsiline nimi, millele kehtestatakse konkreetne struktuur. Andmekogumi faili nimi (koodnimi; füüsiline nimi) on otstarbekas luua nelja tasemelisena:

<avaldaja kood>_<andmekogumi tähis>_<derivatsiooni järjenumber>_<versiooni number>.<laiend>

Sellisest nimest moodustubki andmekogumi unikaalne kood.

Andmete <avaldaja kood> tagab andmekogumite nimede unikaalsuse üle kõigi avaldajate.

<andmekogumi tähis> määrab sama tähendusega andmekogumite seeriad (versioonide jadad).

<derivatsiooni number> peab olema vähemalt kolme kohaline eesnullidega järjenumber ja kirjeldab mitmenda eristruktuuriga samade andmete kogumiga on tegemist. See määrab derivatsioonide järjekorra ja määrab seega derivatsioonide tekkimise ajaloolise järjestuse.

<versiooni number> peab ületäitumise vältimiseks olema vähemalt viiekohaline eesnullidega järjenumber aga sõltuvalt andmete uue versiooni välja andmise sagedusest võib see olla isegi suurem.

<laiend> peab olema kirjeldatud selliselt, et see kirjeldab faili vormingut (N: CSV, JSON, XML, vms.)

Selline andmekogumite unikaalne nimetamine võimaldab andmete kasutajal tuvastada andmete samasust ja vältida samade andmete korduvat töötlust. Selle järgi on võimalik otsida ja läbida ka samade andmetega paralleelseid andmekogumeid

(derivatsioone), et samade andmete kohta käivate täiuslikumate struktuuride alusel võimalusel täiendada juba varem laetud andmeid.

Andmetele ligipääsu hõlbustamiseks tuleb luua hierarhiline andmete sisukord, mis võimaldab andmete poole „käsitsi“ pöördudes leida hõlpsalt vajaliku andmekogumi ja programse pöördumise puhul luua süsteemi kas kõigi või ainult soovitud andmekogumite alamhulga läbimiseks.

Oluline! Pärast andmekogumi derivatsiooni esimese versiooni andmete genereerimist ja avaldamist ei tohi andmekogumi kood-nime ja derivatsiooni numbrit enam muuta. See võimaldab avaandmete kasutajatel lihtsalt aru saada, millised loetavad andmed kuuluvad sama andmekogumi samasse derivatsiooni ja viia omavahel kokku sama andmekogumi sama derivatsiooni erinevad versioonid aga ka sama andmekogumi erinevate derivatsioonide andmed.

Oluline! Andmekogumi versiooni kood-nime ei tohi pärast selle esimest avaldamist enam muuta.

Oluline! Avaldatud andmekogumit ei tohi pärast avaldamist enam muuta st. välistatud peab olema situatsioon, kus sama nime all avalduksid aja möödumisel teistsugused andmed kui nad olid seda esialgselt. See võimaldab avaandete kasutajatel lihtsalt aru saada, milliseid andmeid nad on juba lugenud ja milliseid mitte. Kui toimub andmete parandus tuleb anda välja sama versiooni numbriga (ja seega ka sama nime põhiosaga) andmekogum, mille kood-nimele on lisatud üle alakriipsu paranduse number või tähistada parandus kuidagi teisiti. Peab olema üheselt arusaadav, milliste andmete parandusega on tegemist. Paranduste esitamise loogika peab olema kirjeldatud andmete avaldaja poolt ja avaldatud avaandmete nime standardi kirjelduse koosseisus.

Oluline! Andmekogumit sisaldava faili nimes olevad komponendid peaksid olema korratud andmekogumit sisaldava faili sees so. faili päises kas tervikliku nimena või siis komponentideks lahutatuna. See võimaldab tuvastada faili tähendust ka pärast tema füüsilise nime muutmist.

Vajalikud tegevused:

1. Andmekogumite nimestandardi kinnitamine
2. Faili identifitseerivate tunnuste (koodnime komponentide) lisamine faili päisesse

4.11. Avaandmete sisukorra struktuur

Kõik avaandmed peavad olema leitavad portaali **opendata.riik.ee** kaudu. Paraku on portaali praegune struktuur liiga lame, portaali sisukorrakirje liiga vähe metaandmeid sisaldav ja sobib rohkem andmete visuaalseks otsimiseks kui masinkäsitluseks. Praeguses portaali versioonis ei leia andmete masinkäsitluse kohta üldse infot. Seepärast on mõtet ehitada üles oma avaandmete sisukord (kataloog), mis viitab kõigile meie avaldatud avaandmete kogumitele ja portaali opendata.riik.ee kirjeldada viit sellele sisukorrale. Kui luua vastav kogumite kirjeldus CKAN API abil (docs.ckan.org/en/latest/api/), siis saab portaal opendata.riik.ee need kirjeldused automaatselt endasse hõlmata.

Avaandmete sisukorra kirje päise struktuur peaks sisaldama vähemalt järgmisi andmeid:

1. avaldaja kood
2. andmekogumi tähis
3. derivatsiooni tähis
4. versiooni number

Sisukorra kirjed tuleb sellisel juhul hierarhiliselt grupeerida andmekogumite päise struktuuri alusel:

1. avaldaja kood
2. (avaldaja kood ja) andmekogumi tähis
3. (avaldaja kood, andmekogumi tähis ja) derivatsiooni tähis

Sellisel struktureerimisel sisaldub iga taseme koodi sees ülemuse kood. Võib muidugi kasutada ka ilmutatud viitamist ülemusele.

Iga sellise jaotise kirjelduse sisukorra kirje struktuur peab sisaldama järgmisi andmeid:

1. nimetus
2. selgitus – pikem selgitus taseme kohta.
3. esimese andmekogumi moodustamise aeg
4. viimase (kehtiva) andmekogumi moodustamise aeg
5. viimase (kehtiva) andmekogumi versiooni number
6. sulgemise aeg (aeg millal moodustati viimane andmekogum, millest hiljem enam uusi andmekogumeid ei moodustata)

Avaldaja taseme kirjel on kasutusel andmeelemendid 1-3 ja 6. Kuna siin all on palju erinevaid (erineva kooslusega) andmekogumeid, siis andmeelemente 4 ja 5 ei saa siin kasutada. Nimetuses on kirjeldatud avaldaja juriidiline nimi. Selgituses antakse annotatsioon avaldaja tegevusevaldkonna kohta.

Andmekogumi taseme kirjel on samuti kasutusel andmeelemendid 1-3 ja 6. Seda samal põhjusel, mis avaldaja taseme kirjetel – kuna sama andmekogumi kohta võib olla mitmeid erinevaid derivatsioone siis ei ole üle nende võimalik öelda milline on viimane kehtiv versioon. kuna neid võib olla mitu. Nimetuses esitatakse andmekogumi lühike tähendus. See peab üle kõigi andmekogumite olema unikaalne. Selgituses kirjeldatakse andmekogumi tähendust laiemalt.

Andmekogumi derivatsiooni tasemel on kasutatavad andmeelemendid 1-6.

Kõik sama andmekogumi sama derivatsiooni erinevad versioonid alluvad sisukorra grupeeringu viimasele tasemele.

Kõik nimetused ja selgitused peavad olema vähemalt kahes keeles – eesti ja inglise keeles.

Inim-kasutatavas liideses peab olema võimalik sisukorra otsing otsingu maski alusel – märksõnade ja ajavahemike järgi.

Vaatame nüüd näidet liiklusõnnetuste kohta koostatud avaldatavate andmete näitel.

Olgu meil andmekogumid, mida avaldab Eesti Politsei- ja Piirivalveamet (registri kood 70008747) ja mis sisaldavad liiklusõnnetuste andmeid kuude kaupa. Neid samu andmeid avaldatakse kahes erineva struktuuriga (kaks erinevat derivatsiooni): esimene „Liiklusõnnetused maakondade ja linnade kaupa“ ja „Liiklusõnnetused valdade kaupa“

Avaldaja kood on „70008747“. Andmekogumi kood-nimeks määrame „LKLSNNTS“. Meil on olemas derivatsioonid 001 ja 002. Olgu neist esimeses liiklusõnnetused maakondade ja linnade kaupa ja teises liiklusõnnetused valdade kaupa.

Seega esimese derivatsiooni liiklusõnnetuste andmeid sisaldavate andmekogumite kõik nimed algavad kombinatsiooniga „70008747_LKLSNNTS_001“ ja teise derivatsiooni andmekogumite nimed kombinatsiooniga „70008747_LKLSNNTS_002“. Faili nime moodustamiseks lisatakse derivatsiooni nimele veel versiooni number vormingus „_00001“, „_00002“ jne. Lõppu lisatakse nime laiend, mis näitab andmekogumi versiooni vormingut. Olgu see meie puhul JSON (.JSON).

Selleks, et moodustada sellisel struktuurile sisukord tuleb luua sisukorra grupid avaldajale, andmekogumile ja andmekogumi derivatsioonile:

Tase	**Kood-nimi**	**Nimetus**	**Selgitus**
1	70008747	Politsei- ja Piirivalveamet	Politsei- ja Piirivalveamet tegeleb…
2	70008747_LKLSNNTS	Liiklusõnnetuste andmed	Andmed sisaldavad…
3	70008747_LKLSNNTS_001	Liiklusõnnetuste andmed maakondade ja linnade kaupa	Andmed on grupeeritud…
3	70008747_LKLSNNTS_002	Liiklusõnnetuste andmed valdade kaupa	Andmed on grupeeritud…

Siit on praegu muidugi puudu kõik kuupäevad ja esimeste viimaste versioonide numbri aga antud näite puhule ei oma see tähendust.

Oluline! Selline andmekogumite nimede ja sisukorra struktuur on minimaalne vajaliku sügavusega struktuur. Vastavalt konkreetse andmete avaldaja vajadustele on vahel vajalik vahe-tasemete lisamine igale tasemele. Uusi (lisa)tasemeid ei tohi lisada avaldaja (1.) taseme ette, sest see ajab struktuuri segamine. Tasemed on mõistlik üksteisest eristada eraldussümbolitega (näit. alakriipsuga „ _ “). Lisatavad tasemed ei tohi rikkuda andmekogumite derivatsioonide ja versioonide käesolevas jaotises kirjeldatud järjestuse ja tasemete hierarhilisuse loogikat

<u>Vajalikud tegevused:</u>

1. Andmete sisukorra loomine
2. Sisukorra muutmine vastavalt andmekogumite lisamisele ja kustutamisele
3. Inim-kasutatava liidese ehitamine andmete sisukorra kasutamiseks veebisirvijas.

4.12.Andmete semantiline mõistetavus

Iga avaandmete mistahes andmekogumis oleva andmeelemendi tähendus peab selle kasutajatele olema mõistetav. Selleks peab lisaks andmekogumis olevatele andmetele avaldama ka seal olevate andmeelementide semantika kirjeldused.

Andmeelementide semantika kirjelduste loomiseks on mitu erinevat võimalust

1. Andmeelementide semantika lisatakse iga andmekogumi esimesse kirjesse
2. Andmeelementide semantika lisatakse avaandmete sisukorras oleva andmekogumi derivatsiooni kirjesse (kõik sama derivatsiooni andmekogumid on sama struktuuriga).
3. Andmeelementide semantika lisatakse avaandmete sisukorra andmekogumi (grupi) kirjesse. Sama andmekogumi erinevad derivatsioonid on erineva struktuuriga. Seetõttu kirjeldatakse samas kirjelduses kõikides derivatsioonides esinevate andmeelementide semantika
4. Andmeelementide semantikad kirjeldatakse eraldi sõnastikes ja avaandmete andmekogumid koostatakse selliselt, et need viitavad sõnastikule (JSON-LD, URI, RDF) Kasutada võib avalikke sõnastikke (nt.: http://schema.org)

Vajalikud tegevused:

1. Andmete semantiliste kirjelduste loomine ja sidumine andmekogumitega

4.13.Andmete avaldatuna hoidmise perioodi pikkus

Avaandmete andmekogumi versiooni ei ole mõtet avaldatuna hoida lõpmatuseni. Samas ei tohiks selle avaldamise aeg olla ka liiga lühike. Samuti ei peaks kõikide erinevate andmekogumite versioonide avaldatuna hoidmise aeg olema sama pikk.

Sellest lähtuvalt tuleb iga erineva andmekogu jaoks määrata aeg, mille jooksul peab antud andmekogumi iga versioon olema avaldatud.

Peamiselt on andmekogumi versiooni pikema aja jooksul avaldamise põhjuseks see, et võimaldada uutel kasutajatel alustada mitte „nullist“ so. hetke seisust, vaid saada andmeid ka mõistliku aja kauguselt minevikust ja seega saada oma toode kiiremini tööle ja kasu teenima (nii toote omanikule kui selle kasutajatele). Sellest lähtuvalt peab avaandmete avaldaja otsustama, millise vanusega andmetest võib veel kasu olla, kui me tahaksime teenust pakkuma hakata „täna“. See määrabki konkreetse andmekogumi versioonide avaldatuna hoidmise perioodi pikkuse.

Andmete avaldatuna hoidmise perioodi pikkus ei tohiks üldjuhul olla alla kahe aasta. Selline lähenemine annab andmetele kaheaastase võrdlusbaasi ja võimaldab juba jälgida mingeid trende. Siiski võib sõltuvalt andmekogumi iseloomust olla mõistlik avaldatuna hoidmise aeg nii lühem kui pikem. Samas tuleb siiski iga andmekogumi jaoks määrata see aeg eraldi ja see sõltub andmekogumis olevate andmete iseloomust – kui kiiresti andmed vananevad ja muutuvad mitteasjakohasteks.

Vajalikud tegevused:

1. Analüüsida andmekogumite andmete iseloomu ja määrata igale andmekogumile andmekogumi versioonide avaldatuna hoidmise perioodi pikkus.

4.14.Avaandmete kasutamise jälgimine

Aja jooksul võib sama andmekogumi derivatsioonide hulk kasvada suureks. Seejuures mõnede vanemate derivatsioonide kasutamisest kasutajad loobuvad üldse. On ka võimalik, et mõnesid loodud derivatsioone ei hakatagi kasutama. Sellisest andmekogumitest ja nende derivatsioonidest, mida ei kasutata, on vaja teada saada.

See võimaldab sulgeda mittekasutatavate derivatsioonide andmekogumite versioonide ja võib olla kogu andmekogumi (kui ühtegi selle derivatsiooni pikema aja jooksul ei kasutata) genereerimise.

Samuti on vajalik teada, milliseid andmekogumeid kui palju ja milliseid kanaleid pidi kasutatakse. See võimaldab aru saada andmete kasutatavusest ja produtseerida enam kasutatud valdkondade jaoks rohkem ja võib-olla ka suurema sagedusega andmeid. Kasutuse mõju analüüsi abil saab teha organisatsiooni jaoks olulisi järeldusi, mis aitavad fokuseerida organisatsiooni tegevusi olulistele andmetele, protsessidele jne.

Vajalikud tegevused:

1. Avaandmete kasutusstatistika jälgimismootori loomine
2. Kasutusstatistika andmete kogumine
3. Kasutustatistika andmete analüüs ja sellest tulenevad tegevused:
    a. mitte-kasutatavate andmekogude derivatsioonide sulgemine
    b. mitte-kasutatavate andmekogude sulgemine
    c. uute andmekogumite loomine
    d. olemasolevate andmekogumite uute derivatsioonide loomine
    e. andmekogumite derivatsioonide versioonide genereerimise (loomise) sageduse suurendamine ja vähendamine
    f. andmekogumite kasutuse mõju hindamine, näiteks teabenõuetega mitte tegeletud aja kokkuhoid, andmekvaliteedi tõus vms.

4.15.Avaandmete pidev parendamine

Avaandmete haldamise elutsükkel hõlmab endas nii andmete vigade parandusi kui ka andmekogumite struktuuri edasisi arendusi. Ühelt poolt sattub avaldatavate andmete sisse nii ühe kordseid kui ka regulaarseid vigaseid andmeid ja nende vigade parandamisega tuleb tegeleda. Teiselt poolt nõuab juba eksisteerivate avaandmete kogumite struktuuri muutmist pidevalt muutuv elu – muutvad andmete kooslused, tekkivad uued andmekooslused ja mõningad andmekooslused kaovad. Kõige sellega kaasas püsimine nõuab pidevaid tegevusi.

Kõige vähem avaliku sektori ressurssi koormavaks viisiks vigade leidmisel ja uute arendusvajaduste leidmisel on hoida tagasiside kanal avatuna. Avaandmetest huvitatud kasutajate tagasiside võimaldab efektiivsemalt leida ja parandada vigaseid andmeid ja neid protsesse, mis need vigased andmed tekitasid. Samas on võimalik saada ka vihjeid ja arvamusi selle kohta, millise struktuuriga andmekogumeid oleks vaja luua juurde, milliste olemasolevate andmekogumite struktuure on vaja muuta ja milliste andmekogumite kasutamisest on otstarbekas loobuda.

Enda regulaarsete tegevustega selle kõige saavutamine on oluliselt kallim. Samas ei tohiks ka seda välistada, kuna oma proaktiivse tegevusega avaldatavate andmete koosseisu muutmisel on võimalik kogukondi suunata riiki huvitavatele tegevustele avaandmetega.

Vajalikud tegevused:

1. Kasuta avaandmete kogumi haldamisel mõnda avalikus kasutuses olevat hajusat versioonihaldussüsteemi (nt Github), mis võimaldab veahaldust ja/või

kasutajatel iseseisvalt andmetele parenduse ja arenduse ettepanekuid teha. Alternatiivina paku andmete kasutajatele lihtsal viisil andmete kommenteerimise, täiendamise vms tööriista. Anna tagasisidet.

2. Paranda vastavalt saadud tagasisidele andmeid ja andmete moodustamise protsesse
3. Kaalutle tagasiside kaudu saadud arendusettepanekuid ja arenda selle alusel avaandmete struktuuri
4. Hinda perioodiliselt oma andmete vastavust kriteeriumitele, mis on toodud peatükis 3.
5. Jälgi uute tööandmete lisandumist ja kadumis infosüsteemidest. Tee vajadusel muudatusi avaandmete koosseisus.