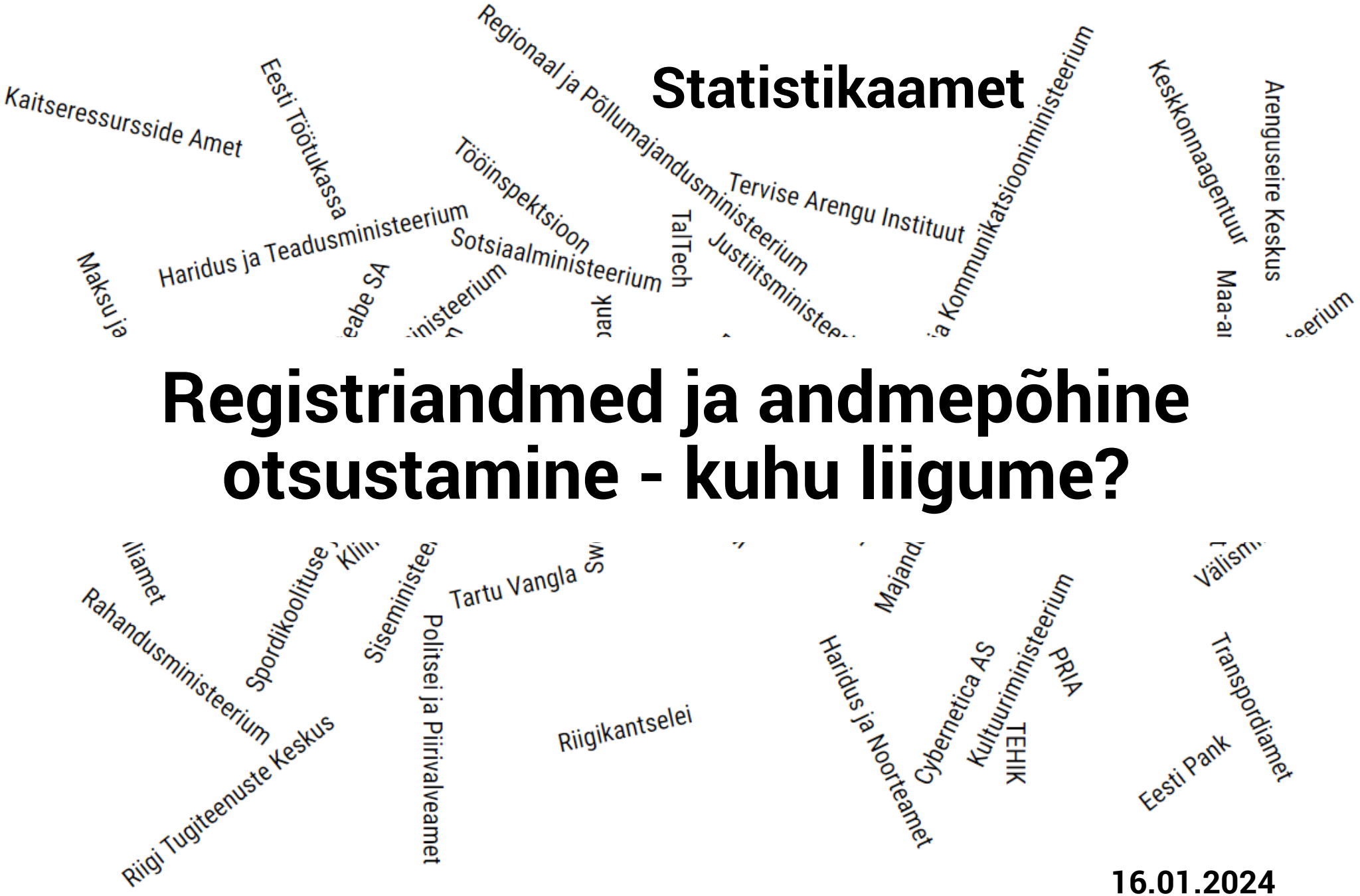


Registriandmed ja andmepõhine otsustamine - kuhu liigume?

EESTI STATISTIKA

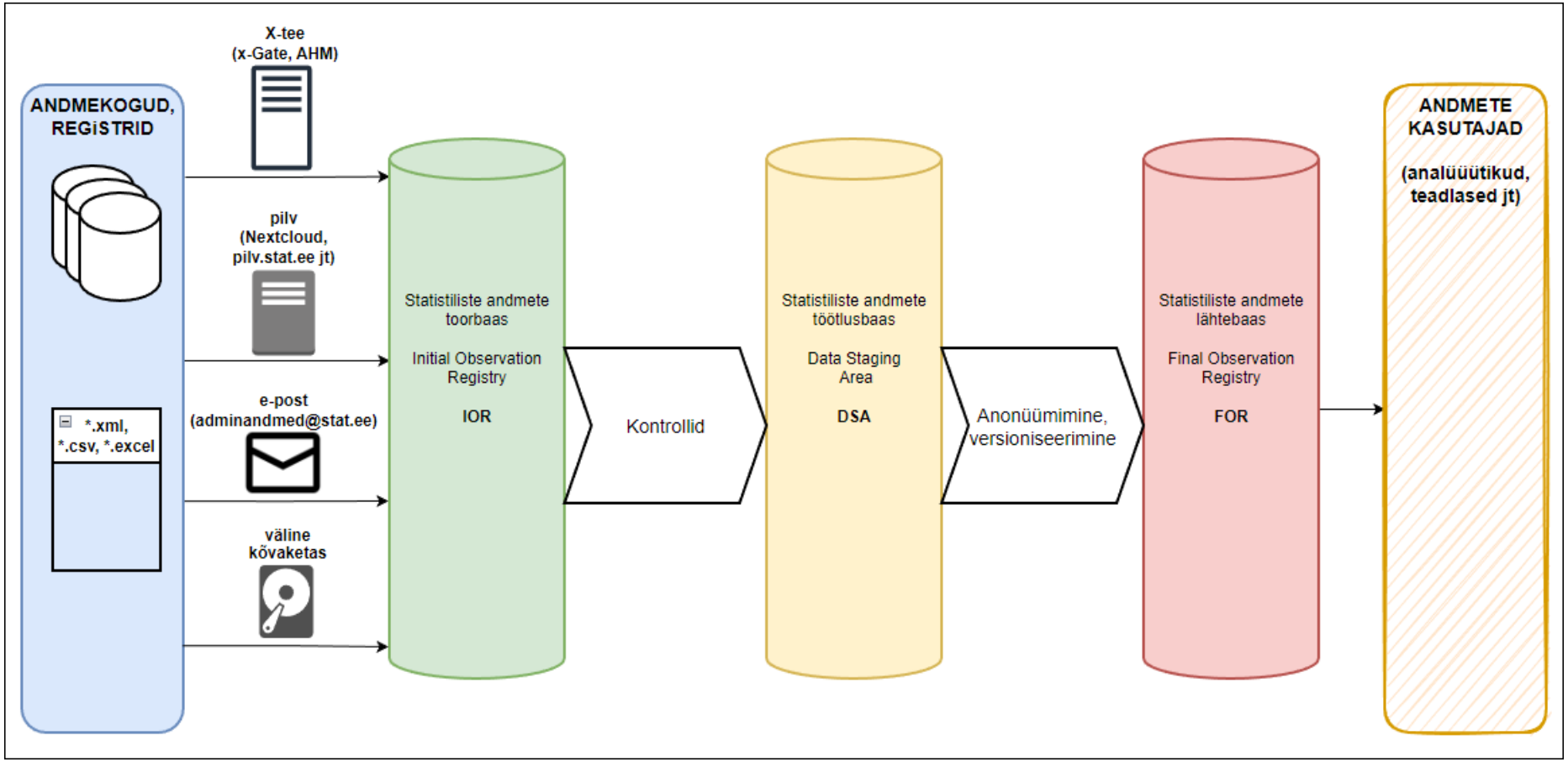
16.01.2024

Andmete hõivamine statistika tegemiseks: mis saab andmetest kui need on Statistikaametisse jõudnud?

Taivo Gross
andmehõive tiimjuht

EESTI
STATISTIKA

Protsessi joonis



Andmete laekumisel esinevad probleemid

- Andmestruktuur ei vasta kokkulepitule
- Topeltkirjed
- Vale kirjete arv
- Probleemid kuupäevadega
- Probleemid aadressidega
- Tähemärkide probleemid
- Andmeedastuse automaatika probleem
- Sisulised küsimused andmete kohta
- Korrektse andmeedastuse hilinemine
- Muud probleemid andmetes

Andmestruktuur ei vasta kokkulepitule

Välja nimetus	Välja formaat	Välja selgitus
KOOD	String	Isikukood
SYNNIKP	Date	Sünniaeg, kui puudub Eesti isikukood
VM_KOOD	String	Väljamaksja kood
AASTA	Number	Väljamakse tegemise aasta
KUU	Number	Väljamakse tegemise kuu
RIIK	String	Riigi kood
…		

Näited:

- väljade arv erineb kokkulepitust – kas on mõni väli puudu või on välju rohkem kui kokkuleppes;
- välja nimetus erineb kokkulepitust – nt 'KOOD' asemel 'IK';
- välja formaat erineb kokkulepitust – nt 'KUU' on 'jaanuar' kuigi kokkuleppes on kuu numbriline väli;
- …

Topeltkirjed

- Topeltkirjete vältimiseks lepitakse andmeedastuslepingus enamasti kokku andmete struktuur ja väljavõtte tingimused selliselt, et oleks tagatud iga kirje unikaalsus.
- Duplikaatideks on andmetabeli read, millel on sama unikaalne võti. Unikaalne võti võib koosneda ühest (nt. ISIKUKOOD) või mitmest väljast (nt ISIKUKOOD, PERIOOD või ISIKUKOOD, PERIOOD, TULU_LIIK).
- Mõnes andmestikus puudub kirjetel unikaalne võti ja sel juhul topeltkirjete kontrolli andmehõive käigus ei teostata.

Ettevõtja nimi	Registrikood	Kehtivuse algus	Kehtivuse lõpp	Kehtiv
Osaühing Testisik	10234511	14.08.2021	14.08.2031	Jah
Osaühing Testisik	10234511	14.08. 2021	14.08.2031	Jah

Vale kirjete arv

Põhjused:

- Väljavõtte tingimusi pole korrektselt rakendatud
- Andmeesitaja laekuvad andmed mõnelt teiselt andmekogult
- Võimalikud tehnilised probleemid

Lahenduseks on, et koos andmetega saadab andmeesitaja info ka kirjete arvu kohta

Probleemid kuupäevadega

Näited:

- Kokku lepitud kuupäeva formaat näiteks '31.01.2022', kuid laekunud andmetes on kasutatud teistsugust formaati, nt '2022-01-31T00:00:000' vms
- Ebatõenäolised kuupäevad, nt sünnikuupäevana 1202-04-01, või ka olukord kus kuupäeva asemel esitatakse vaid aastanumber.
- Ühe muutuja lõikes on väärtused erinevas formaadis (st osad tunnused kuupäevad ja osad ainult aasta number)

Eelistatud lahendusena soovime põhjustele selgitust andmeesitajalt ning võimalusel parandatud andmed.

Probleemid aadressidega

- ADS-iga liidestatuse olulisus
- Ka ADS-iga aeg-ajalt tõrked, mittekvaliteetne sisend
- ADS-i kujul saabunud aadressobjektide lisatöötluse vajadus
- Aadressiandmete kvaliteedi parandamine ja lisatöötlus on väga ajamahukas

Tähemärkide probleemid

Andmete edastamisel probleemid tähemärkide kodeeringuga

Näiteks:

„M˘GI“, peaks olema „MÄGI“,

„PREOBRA˘ENSKI“, peaks olema „PREOBRAŽENSKI“.

Võimalikud põhjused:

UTF-8 formaadis salvestatud andmed on eksporditud ANSI formaati

Andmeedastuse automaatika probleem

Kas ikka 100% veavaba? Hilinemise põhjustest:

- Automaatsete andmete edastust ei toimunud
- Andmeesitajast mittesõltuvad asjaolud
- Tehnilised väljakutsed. X-tee kaudu sama andmeedastuse käivitamine.
- Andmete edastusel *time out* viga

Sisulised küsimused andmete kohta

Ilmnevad etapis, mil analüütikud andmeid analüüsivad:

- Kirjete arv väheneb/suureneb oluliselt
- Mõnel kirjel sündmuse alguskuupäev hilisem kui lõppkuupäev
- Kuupäev ebatõenäolises ajavahemikus

Korrektse andmeedastuse hilinemine

- Andmehõive kuupäevadest kinnipidamise olulisus
- Andmestike hilinemise mõju edasisele andmetöötlusele
- Andmete hilinemine võib mõjutada avaldamiskalendrit

Muud probleemid andmetes

- Andmestikus tühjade väärtuste asemel punkt
- Andmete väljavõtt valede tingimustega
- Andmestik ei laekunud terviklikult
- Andmekogud märkasid ise peale andmete esitamist puudujääke kvaliteedis ning soovisid ise andmed uuesti esitada

Tänan!

Taivo Gross
andmehõive tiimjuht
taivo.gross@stat.ee

EESTI STATISTIKA
www.stat.ee
Tatari 51, 10134 Tallinn