Labradata LabDW

Document author: anna.hammais@tyks.fi

Raakadata

labdw_data

Raakadataa. Lähes joka sarakkeessa koodi, joka täytyy kääntää selitteeksi käyttäen kyseiselle sarakkeelle määritettyä koodistoa. Raakadata poimitaan viikottain keskiviikkoisin biopankki-siirtokansioon.

labdw_columns

Sarakkeille määritetyt koodistot ja sarakkeiden nimien selitteet löytyvät tiedostosta labdw_columns. Samalle sarakkeelle on usein määritelty kaksi koodistoa, joissa onneksi ei lähes koskaan ole päällekkäisiä koodiarvoja.

labdw_codetables

Koodiarvojen selitteet kullekin koodistolle löytyvät tiedostosta labdw_codetables. Sarake codetable kertoo koodiston, code on koodi ja text on selite.

Datan avaimet

Raakadatassa ei ole pääavainta. testid-sarake (Tutkimusriviavain) on eräänlainen tutkimuspaketin tunniste, ja tutkimuspakettiin kuuluu monta eri tutkimusta. etlstamp (stage_hadoop-kannassa “dt”, aurian näkymissä “labdw_biopankki_poiminta_pvm”) kuvaa ilmeisesti ajankohtaa, jolloin tieto on tuotu labdw-tietokantaan MultiLabista. Pääavaimen tapaisena voi käyttää yhdistelmää {testid, test}, eli tietyn testipaketin tiettyä testiä, ja etl-aikaleiman perusteella päätellään, että kyseessä on saman testin uusi versio, eli tieto päivitetään vanhan päälle.

etlstamp-arvo on virallisesti lähdekannassa merkkijono, joka esittää aikaleima muodossa (Postgres-tyyliin) YYYYMMDDHH24MISS (Javaksi yyyyMMddHHmmSS). Kuitenkin joidenkin merkkijonojen alussa on sana “rerun”, jonka jälkeen numero-osa tulee. Muita merkkijonoja ei esiinny ainakaan tällä hetkellä (2016-09-08).

Väliaikaiseen käyttöön tehdyssä puhdistetussa labradatassa (esim. lab_temp.temp_lab_vsshp_join) täydelliset rividuplikaatit on poistettu ja etlstamp-sarakkeen perusteella riveistä on säilytetty vain uusin. Silti löytyy tekstimuotoisia vastauksia, joilla on sama Tutkimusriviavain ja sama Tutkimus ja sama etlstamp/dt (eli ovat poimiutuneet samassa poiminnassa, eli olleet samaan aikaan olemassa lähdejärjestelmässä), mutta eri tulos. Näille ei voida oikein mitään, onneksi näitä on vain yhtä testiä eikä tulos ole numeerinen

Vuosi 2004

```
select extract(year from datetime3), count(*)
from stage_labdw.labdw_transform
where source_table = 'labdw_tutkimus_2004'
group by extract(year from datetime3)
order by extract(year from datetime3);

-- 1997 1
-- 1999 2
-- 2000 6980
```

```
-- 2001 506401
-- 2003 651
-- 2004 788270
--  28
```