Rakennusten osoitetietojen haltuunotto

Jukka Rahkonen
http://latuviitta.org

Viimeksi muokattu 21. kesäkuuta 2016

Väestötietojärjestelmästä poimitut rakennusta osoitetiedot on saatavilla osoitteesta https://www.avoindata.fi/data/dataset/rakennusten-osoitetiedot-koko-suomi

Aineisto on saatavilla yhdeksänätoista tekstitiedostona, jotka sisältävät tämän näköistä tietoa:

```
102183955L;091;01;1;6684510;389337;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183956M;091;01;1;6684504;389359;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183957N;091;01;1;6684491;389381;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183958P;091;01;1;6684274;389350;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183959R;091;01;1;6684308;389344;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183960S;091;01;1;6684320;389377;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183961T;091;01;1;6684348;389359;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183962U;091;01;1;6684328;389316;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183963V;091;01;1;6684351;389330;1;Pohjantähdentie;Polstjärnevägen;2;00740
102183964W;091;01;1;6684357;389306;1;Pohjantähdentie;Polstjärnevägen;2;00740
```

Tekstitiedoston rakenne on kuvattu dokumentissa https://www.avoindata.fi/dataset/cf9208dc-63a9-44a2-9312-bbd2c3952596/resource/d5f9dda4-6a1a-4b27-a87c-41b7bed0160e/download/Rakennusten-osoitetiedot-Avoin-data-kuvaus-tiedoista-2016-04-20.pdf

Kentän nimi	Tyyppi ja pituus	Selitys
Rakennustunnus	char(20)	Pysyvä rakennustunnus. 10 merkkiä pitkä aakkosnumeerinen tunnus, jonka ensimmäinen merkki on ”1” ja viimeinen merkki Modulo-31 säännöllä laskettu tarkistusmerkki, joka on joko numero tai iso kirjain. Tarkistusmerkin laskentasääntö ja käytössä olevat tarkistusmerkit ovat samat kuin henkilötunnuksen.
Sijaintikunta	char(3)	Kunta, jossa rakennus sijaitsee. Kuntanumero, sisältää etunollat.
Maakunta	char(2)	Maakunnan numero, sisältää etunollat.

Käyttötarkoitus	char(1)	Rakennuksen käyttötarkoituksen mukaisesti, johdetut koodiarvot: **1** = Asuin- tai toimitilarakennus (rakennuksen käyttötarkoitus väliltä 001-599) **2** = Tuotanto- tai muu rakennus (rakennuksen käyttötarkoitus väliltä 600-999) **0** = Rakennuksen käyttötarkoituksesta ei tietoa (puuttuu tai 0)
Pohjoiskoordinaatti	number(7)	Rakennuksen keskipisteen pohjoiskoordinaatti. Koordinaatisto ETRS-TM35FIN
Itäkoordinaatti	number(6)	Rakennuksen keskipisteen itäkoordinaatti. Koordinaatisto ETRS-TM35FIN
Osoitenumero	number(1)	Moniosoitteisen rakennuksen osoitteen järjestysnumero.
Kadunnimi suomeksi	char(100)	Kadunnimi suomeksi
Kadunnimi ruotsiksi	char(100)	Kadunnimi ruotsiksi
Katunumero	char(7)	Katunumero
Postinumero	char(5)	Postinumeroalueen tunnus, etunollat.

Rakennusosoiteaineiston haltuunotto GDAL-ohjelmilla

Tavoite:

1. Kootaan aineisto yhdeksi koko maan kattavaksi aineistoksi
2. Muunnetaan aineisto paikkatietomuotoon muuntamalla rakennusten itä- ja pohjoiskoordinaatit pistemäisiksi geometrioiksi ETRS-TM35FIN-koordinaattijärjestelmään
3. Johonkin sellaiseen tiedostomuotoon, joka voidaan avata tehokkaasti paikkatieto-ohjelmistoilla

Tunnistetut ongelmat:

- Testitiedosto ei sisällä tietoa kentän tietotyypistä. Esimerkiksi kuntanumerot ja postinumerot voivat alkaa etunollilla, minkä takia on varmistetttava, että ne tulkitaan tekstiksi eikä numeroiksi. Usein tämä erottelu sisältyy tekstiedostoon siten, että tekstiksi tulkittavat kentät on laitettu lainausmerkkien sisään:
  123;091 → numerot 123 ja 91
  ”123”;”091” → tekstit 123 ja 091
- Nimissä käytettyä merkistökoodausta ei ole kerrottu. Se on todennäköisesti ISO 8859-1 eli Latin1. Tämä oletus toimii suomen- ja ruotsinkielisillä nimillä, ja saamenkielisiä nimiä aineistossa ei olekaan.

Suoritus:

Aineisto on tekstimuotoisena ja tietueiden kentät on erotettu puolipisteellä, joten sitä voidaan käsitellä GDAL:in CSV-ajurilla (Comma separated values)
http://www.gdal.org/drv_csv.html

Suoritukseen kuuluu seuraavat vaiheet:

1. Kenttien tietotyyppien määrittely
2. Kenttien nimien määrittely
3. Erillisten maakunnittaisten aineistojen yhdistäminen
4. Merkistökoodauksen muunnos UTF-8:aan
5. Aineiston muunnos paikkatietomuotoon

Kenttien tietotyyppien määrittely

Tietotyypit voidaan määritellä kirjoittamalla ne pilkuilla erotettuna tekstitiedostoon, jonka nimeksi annetaan tekstiedoston nimen pääosa ja tarkentimeksi .csvt. Yllä olevan taulukon mukainen määrittely saadaan aikaan tällaisella tekstitiedostolla, joka tallennetaan nimellä ”osoitteet.csvt” odottamaan tulevaa tarvetta:

```
"String","String","String","String","CoordY","CoordX","Integer",
"String","String","String","String"
```

Tiedosto tallennetaan yhdelle riville ilman rivinvaihtoja. Huomaa erityiset tietotyypit CoordX ja CoordY, joilla voidaan määritellä itä- ja pohjoiskoordinaatit. Huomaa myös, että tässä määrittelyssä X on aina itäkoordinaatti tai pituusaste ja Y on pohjoiskoordinaatti tai leveysaste riippumatta siitä, mikä koordinaattijärjestelmä on käytössä.

Kenttien nimien määrittely

GDAL tunnistaa kenttien nimet automaattisesti, jos ne tallennetaan tekstitiedoston ensimmäiselle riville. Rakennusosoiteaineisto on kuitenkin jaetty 19 erilliseen tekstitiedostoon, mistä aiheutuu pieniä ongelmia. Niistä selvitään esimerkiksi niin, että tallennetaan kenttien nimet omaan tekstitiedostoon, jossa nimet ovat yhdellä rivillä ja rivin loppuun tulee rivinvaihto. Tallennetaan tiedosto esimerkiksi nimellä ”kenttien_nimet.csv”

```
rakennustunnus;sijaintikunta;sijaintimaakunta;rakennustyyppi;
CoordY;CoordX;osoitenumero;katunimi_fi;katunimi_se;katunumero;
postinumero
```

Aineistojen yhdistäminen

Kenttien nimet ja 19 osoitetiedostoa voidaan yhdistään Windows:in Copy-komennolla.

```
copy kenttien_nimet.csv+*.OPT osoitteet_latin1.csv
```

Tuloksena on tiedosto ”osoitteet_latin1.csv”, johon on kirjoitettu ensimmäiseksi kenttien nimet ja sen jälkeen kaikki .OPT-päätteiset rakennusoitetiedot. Kenttien nimet sisältävän tiedoston rivinvaihtomerkin tärkeys selviää tässä: jos rivinvaihto puuttuu, niin ensimmäiselle riville tulee sekä nimet että ensimmäinen osoitetietorivi.
Kysymys: Miksei vain yhdistetä nimitiedostoja ja kirjoiteta kenttien otsikot tekstinkäsittelyohjellla ensimmäiseksi riviksi?
Vastaus: Sitä voi ihan hyvin yrittää, mutta yhdistetyssä tekstitiedostossa on 3,5 miljoonaa riviä eivätkä kaikki tekstinkäsittelyohjelmat pysty avaamaan sitä.

Merkistökoodauksen muunnos

GDAL:in CSV-ajuri olettaa, että tekstissä käytetään UTF-8-merkistökoodausta. Muunnos Latin1-koodauksesta UTF-8-koodaukseen voidaan tehdä tekstinkäsittelyohjelmalla, esimerkiksi Notepad++:lla, tai komentoriviltä iconv-ohjelmalla. Iconv-ohjelman asentamista Windows:lle ei käsitellä tässä ohjeessa. Jos iconv on käytettävissä, niin muunnoskomento on:

```
iconv -f latin1 -t UTF-8 osoitteet_latin1.csv >osoitteet.csv
```

Muunnos paikkatietomuotoon

Nyt rakennuspisteaineisto on valmisteltu niin pitkälle, että se voidaan muuntaa ogr2ogr-ohjelmalla johonkin käyttökelpoisempaan muotoon. Kuten aina, ennen muunnosta kannattaa vilkaista ogrinfo-ohjelmalla, että kaikki näyttää olevan kunnossa:

```
ogrinfo -dialect sqlite -sql "select * from osoitteet limit 1"
osoitteet.csv

INFO: Open of `osoitteet.csv'
      using driver `CSV' successful.

Layer name: SELECT
Geometry: Point
Feature Count: 1
Extent: (421305.000000, 6718713.000000) - (421305.000000,
6718713.000000)
Layer SRS WKT:
(unknown)
Geometry Column = GEOMETRY
rakennustunnus: String (0.0)
sijaintikunta: String (0.0)
```

```
sijaintimaakunta: String (0.0)
rakennustyyppi: Integer (0.0)
CoordY: Real (0.0)
CoordX: Real (0.0)
osoitenumero: Integer (0.0)
katunimi_fi: String (0.0)
katunimi_se: String (0.0)
katunumero: String (0.0)
postinumero: String (0.0)
OGRFeature(SELECT):0
  rakennustunnus (String) = 100220387P
  sijaintikunta (String) = 018
  sijaintimaakunta (String) = 01
  rakennustyyppi (Integer) = 1
  CoordY (Real) = 6718713
  CoordX (Real) = 421305
  osoitenumero (Integer) = 1
  katunimi_fi (String) = Tarkintie
  katunimi_se (String) =
  katunumero (String) = 129
  postinumero (String) = 07530
  POINT (421305 6718713)
```

Tulos näyttää varsin hyvältä, joten tehdään seuraavaksi varsinainen muunnos. SpatiaLite tai GeoPackage ovat erittäin hyviä tiedostomuotoja tällekin aineistolle; valitaan tällä kertaa SpatiaLite. Komento, joka tekee muunnoksen ja kertoo myös, että rakennusten koordinaatit ovat ETRS-TM35FIN-järjestelmän mukaisia (koodi EPSG:3067) on seuraava:

```
ogr2ogr -f sqlite -dsco spatialite=yes -a_srs epsg:3067
osoitteet.sqlite osoitteet.csv
```

Lopputulosta voidaan tarkastella karttaohjelmalla. Rakennustunnukset, kuntanumerot, katunimet jne. näyttävät siirtyneen juuri oikein, mukaan lukien tunnuksiin ja postinumeroihin kuuluvat etunollat. Jotkin osoitepisteet näyttävät osuvan Suomen ulkopuolelle, mutta tähän näyttää olevan syynä yksinkertaisesti se, että koordinaatit ovat väärin lähtöaineistossa. Tässä esimerkkirivi, joka on poimittu suoraan ladatusta .OPT-tiedostosta:

```
101382787F;178;10;1;6873007;873467;1;Kalajärventie;;383;51980
```

Extra: Luo indeksejä

Koska osoitteet nyt ovat SpatiaLite-tietokannassa, niin saattaa olla järkevää tehdä indeksit niille kentille, joiden mukaan tullaan tekemään hakuja. Jokaisesta indeksistä maksetaan tietokannan koon suurenemisen muodossa, joten summamutikassa ei kannata indeksoida kaikkia kenttiä. Indeksien luominen onnistuu erimerkiksi ogrinfo-ohjelmalla.

```
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "rakennustunnus_idx" ON "osoitteet"
("rakennustunnus")"
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "sijaintikunta_idx" ON "osoitteet"
("sijaintikunta")"
```

```
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "sijaintimaakunta_idx" ON
"osoitteet" ("sijaintimaakunta")"
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "rakennustyyppi_idx" ON "osoitteet"
("rakennustyyppi")"
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "katunimi_fi_idx" ON "osoitteet"
("katunimi_fi")"
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "katunimi_sv_idx" ON "osoitteet"
("katunimi_sv")"
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "katunumero_idx" ON "osoitteet"
("katunumero")"
ogrinfo osoitteet.sqlite -sql "CREATE INDEX "postinumero_idx" ON "osoitteet"
("postinumero")"
```

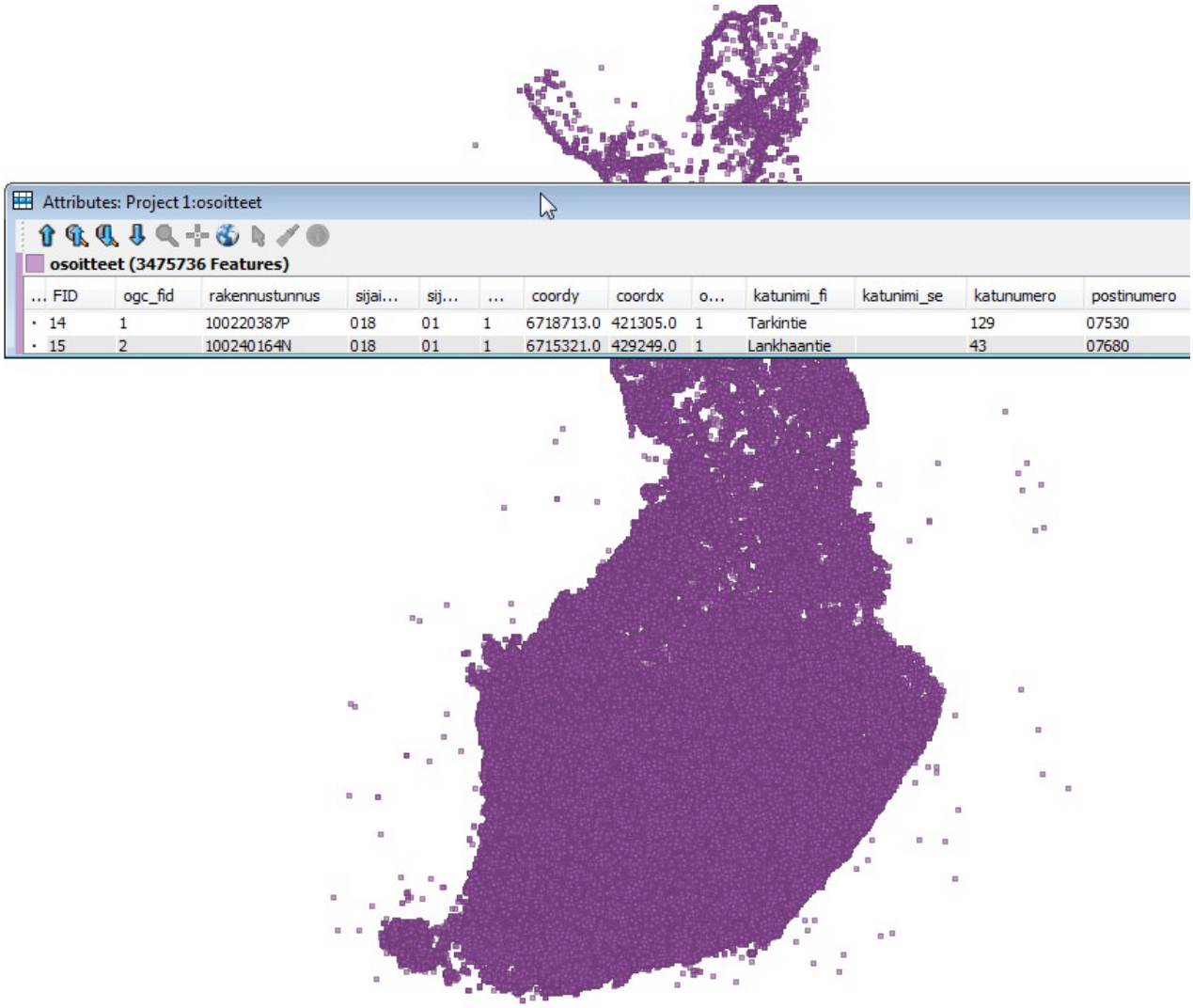
Indeksien vaikutus tietokannan kokoon

Tässä taulukossa on osoitetietokannan koko tavuina ilman indeksejä (0) sekä jokaisen yllä esitetyn indeksin luomisen jälkeen.

```
0:  650047488
1:  717424640
2:  759928832
3:  798740480
4:  832007168
5:  901510144
6:  940428288
7:  978386944
8: 1027979264
```

Kysymys: Kestääköhän tämä aineiston haltuunotto kauan?
Vastaus: Eipä oikeastaan näiden mitattujen tulosten perusteella:

Kenttien nimien ja OPT-tiedostojen yhdistäminen: **20 sekuntia**
Iconv-muunnos UTF-8:aan: **1 minuutti 40 sekuntia**
SpatiaLite-tietokannan luominen: **3 minuuttia**
Indeksien luominen: **3 minuuttia**
Yhteensä: **8 minuuttia**


Attributes: Project 1:osoitteet
osoitteet (3475736 Features)
... FID ogc_fid rakennustunnus sijai... sij... ... coordy coordx o... katunimi_fi katunimi_se katunumero postinumero
14 1 100220387P 018 01 1 6718713.0 421305.0 1 Tarkintie 129 07530
15 2 100240164N 018 01 1 6715321.0 429249.0 1 Lankhaantie 43 07680