

Maanmittauslaitoksen vuoden 2011 XML-muotoisten kuntaraja-aineistojen haltuunotto

Tässä selostettu menetelmä toimii Maanmittauslaitoksen 9. marraskuuta 2011 päivitettyillä XML-muotoisilla kuntajakoaineistoilla, jotka on ladattu osoitteesta <http://www.maanmittauslaitos.fi/ilmaisetaineistot>

Inspire-yhteensopiva aineisto

Kuntajakoaineistot on tehty Inspire-direktiivin aluajakotietotuotteen määrittelyjen mukaisesti, mistä saa lisätietoja näistä asiakirjoista:

http://xml.nls.fi/Kuntajako/Asiakasdokumentaatio/Skeemakuvaus/INSPIRE_DataSpecification_AU_v3.0.1.pdf
http://xml.nls.fi/Kuntajako/Asiakasdokumentaatio/Skeemakuvaus/NLSF_AU_schema_documentation.html

Skeematiedosto, jota kuvaa XML-tiedoston rakenteen ja joka on välttämätön sen täydelliseen tulkitsemiseen on ladattava erikseen osoitteesta

<http://xml.nls.fi/inspire/2010-04/AdministrativeUnits.xsd>

Tavalliselle valmisohjelmistojen käyttäjälle vaikea aineisto

XML-muotoisia kuntajakoaineistoja ei pysty avaamaan suoraan millään avoimen lähdekoodin paikkatieto-ohjelmistolla (Quantum GIS, gvSIG, Kosmo, OpenJUMP, uDig ym.) sen enempää kuin valtaosalla kaupallisiakaan ohjelmistoja (ArcView, MapInfo). Tämä johtuu siitä, että valtaosa maailman paikkatietosovelluksista on tehty tukemaan varsin yksinkertaista tietomallia, jossa kohteet ovat yksinkertaisia kohteita, ja niiden ominaisuustiedot esiintyvät vain kertalleen. Samannäköinen tilanne siis kuin jos ottaa taulukkolaskentaohjelmasta yhden rivit. OGC-standardeista tätä tapausta kuvaa "Compliance level SF-0" of "GML 3.1.1 simple features profile - OGC(R) 06-049r1" : http://portal.opengeospatial.org/files/?artifact_id=15201

Inspire-direktiivin aluejakotietotuote ei ole "Simple Features" -määrittelyn mukainen, joten sitä on mahdotonta avata voin yksinkertaista skeemaa tukevilla ohjelmilla. Onneksi tuon rakenteen mukaisesta tiedosta on mahdollista kaivaa esille yksinkertaisen mallin mukaiset viiva- ja aluetasot. Muunnoksessa hukkuu osa tietosisällöstä, mutta on paljon parempi saada vakiosovelluksiin valtaosa hyödyllisestä tiedosta kuin jäädä kokonaan ilman.

XML-kuntajakoaineiston tunnistetuista ongelmista pienin on se, että tiedostoihin sisältyy kaksi erillistä karttatasoa. Toisella tasolla on kaikki aluejaon viivat viivamaisena tasona (LineString) ja toisella tasolla aluejaon kohteet aluekohteina (MultiPolygon). Hankalampi ongelma on se, että aluetasolle kuntien toisten kuntien sisällä sijaitsevia alueita eli enklaveja hallitaan niille annettulla enclave-roolilla (Role). Vain normaalia perus-GML:ää ymmärtävät ohjelmat eivät tiedä, mitä Role merkitsee. Parhaassa tapauksessa ne hyppäävät Role-elementeistä yli, pahimmassa tapauksessa ne lopettavat XML-tiedoston jäsentämisen kokonaan, antavat virheilmoituksen ja käyttäjä jää kokonaan ilman paikkatietoja. Toinen hankaluus on StringList-tietotyyppiä käyttävät ominaisuustiedot. Tämä tietotyyppi tarkoittaa sitä, että yhdelle ominaisuustiedolle annetaan luettelomuodossa monta eri arvoa, erimerkiksi paikkakunnan nimi suomeksi ja ruotsiksi. Valtaosa paikkatieto-ohjelmistoista tukee vain "yksi ominaisuustieto – korkeintaan yksi arvo" -periaatetta.

Inspire-XML on tarkoitettu Inspire-yhteensopivien järjestelmien väliseen paikkatietojen vaihtoon ja siihen tarkoitukseen se varmasti sopii. Tosin siinäkin kehitys taitaa mennä niin päin, että kun tietomalli on ensin määritelty, niin sitten tehdään vaikka väkisin ohjelmia, jotka voivat jotenkin hyödyntää sitä. Tavallisten paikkatieto-ohjelmien käyttäjiä ei määrittelyjä tehtäessä ole ajateltu ollenkaan, sillä muussa tapauksessa olisi varmaankin annettu ohjeet tai mieluummin valmis

ilmainen sovellus aineiston oleellisten osien muuntamiseksi Simple Features -mallin mukaiseksi.

MML:n kuntajako-XML:n haltuunotto ogr2ogr-ohjelmistolla

Valmisohjelmistojen käyttäjän onneksi kuntajako-XML on mahdollista muuntaa helpommin käsiteltävään muotoon GDAL-ohjelmakirjastoon kuuluvalla ogr2ogr-ohjelmalla. GDAL on ilmainen, avoimen lähdekoodin ohjelma, jota voi käyttää vapaasti mihin tahansa tarkoitukseen, myös kaupallisiin.

Ogr2ogr-ohjelma ei tue täysin aluejakoaineistojen tietomallia, se ei esimerkiksi tiedä, miten enklaaveiksi merkityjä alueita kuuluisi käsitellä. Se pystyy kuitenkin muodostamaan XML-tiedostosta normaalin viivamaisen ja aluemaisen tason ja yhdistämään niiden geometrioihin niille kuuluvat ominaisuustiedot. Muunnos voidaan tehdä mihin tahansa tiedostomuotoon, johon GDAL pystyy kirjoittamaan. Näitä ovat kaikki tiedostomuodot, joilla on arvo Creation – Yes taulukossa http://gdal.org/ogr/ogr_formats.html ja niihin kuuluu mm. ESRI shapefile, MapInfo tab ja mid/mif sekä Microstation dgn. Ogr2ogr voi kirjoittaa myös suoraan useisiin tietokantoihin, kuten Oracle Spatial, MS SQL ja PostgreSQL.

Tässä ohjeessa neuvotaan muunnoksen tekeminen Spatialite-tietokantaan sekä ESRI shapefile- ja MapInfo tab -tiedostomuotoihin. Spatialiteä on mukana siitä syystä, että tietokantana, siihen voidaan kirjoittaa molemmat alkuperäiseen aineistoon kuuluvat karttatasot. Lisäksi Shapefile-ajuri tukee Maanmittauslaitoksen XML-aineistoissa esiintyvää StringList-tietotyyppiä, jota shapefile ja MapInfo tab eivät tue.

Shapefile-muodon käytössä on muitakin rajoituksia kuin puuttuva tuki StringList-tietotyyppille. Yksi shapefile voi sisältää vain yhden tyyppisiä geometrioita, joten viivoista ja alueista pitää kirjoittaa omat shapefilensä. Lisäksi ominaisuustietojen nimissä ei shapefilessä voi olla kuin korkeintaan 10 merkkiä, joten jotkut alkuperäisen aineiston eräät pitkät nimet, kuten ”LocalisedCharacterString”, katkeavat hankalasti luettavaan muotoon. Spatialitelle ei ole ongelmia myöskään pitkien ominaisuustietokenttien nimien takia.

MapInfo tab voisi periaatteessa sisältää useita karttatasoja, mutta GDAL:n ”MapInfo file” -ajuri ei tätä tue. Siitä syystä ogr2ogr-ohjelmalla on tehtävä erilliset viiva- ja aluetiedostot myös MapInfo-muotoon muunnettaessa.

Oikean GDAL-version hankkiminen Windows-koneelle

Marraskuussa 2011 julkaistut viralliset GDAL-version eivät pysty muuntamaan Maanmittauslaitoksen XML-tiedostoista muuta kuin viivatason. Kuitenkin kun ohjelman kehittäjille toimitettiin esimerkkiaineisto, niin myös aluetason muuntamisen mahdollistavat muutokset tehtiin ohjelmistoon alle viikossa. GDAL:in kehitysversiolla pystytään siis nyt muuntamaan sekä viiva- että aluetasot.

Windows-käyttäjä saa GDAL-kehitysversion käyttöönsä helpoiten lataamalla MSI-asennusohjelman osoitteesta <http://www.gisinternals.com/sdk/> Esimerkkimuunnos käytettiin käyttämällä asennuspakettia *gdal-19dev-1600-core.msi* joka ladattiin sivulta <http://www.gisinternals.com/sdk/PackageList.aspx?file=release-1600-gdal-mapserver.zip>.

XML-tiedoston ominaisuuksien tutkiminen ja tulkinta

Ennen minkään muunnoksen tekemistä ogr2ogr-ohjelmalla kannattaa tutkia muunnettavaa aineistoa ogrinfo-ohjelmalla

```
D:\temp>ogrinfo SuomenKuntajako_2011_4500k.xml
ERROR 4: Failed to open SuomenKuntajako_2011_4500k.xsd to read.
Had to open data source read-only.
INFO: Open of `SuomenKuntajako_2011_4500k.xml'
      using driver `GML' successful.
1: AdministrativeUnit (Multi Polygon)
2: AdministrativeBoundary (Line String)
```

Listauksen tulkinta

- ERROR 4 kertoo, että tässä tapauksessa samasta hakemistosta missä SuomenKuntajako_2011_4500k.xml -tiedosto on, ei löytynyt sitä vastaavaa .xsd -pääteistä skeematiedostoa. Virhe ei tässä tapauksessa ole vakava, ogr2ogr selviää tehtävästään myös ilman xsd-tiedostoa. Itse asiassa se ei käytä skeematiedostoa, vaikka se löytyisikin, koska sen kaikkia Inspire-erikoisuuksia ei voida muuntaa ”Simple Feature” -tietomalliin, jota GDAL/OGR sisäisesti käyttää.
- INFO kertoo, että tiedostosta on löytynyt kaksi karttatasoa: aluekohteita sisältävä AdministrativeUnit, ja viivoja sisältävä AdministrativeBoundary.

Tasoista voidaan pyytää tarkempia tietoja antamalla tiedoston nimi ja tason nimi peräkkäin. Summary only (-so) parametriä on järkevää käyttää ainakin aluksi, koska muuten ogrinfo listaa tiedot myös tason jokaisesta kohteesta, mukaanlukien sen geometrian tekstimuotoisena esityksenä.

```
D:\temp>ogrinfo -so SuomenKuntajako_2011_4500k.xml AdministrativeUnit

Layer name: AdministrativeUnit
Geometry: Multi Polygon
Feature Count: 362
Extent: (61686.154000, 6605838.532000) - (733893.585000, 7778595.169000)
Layer SRS WKT:
PROJCS["ETRS89 / TM35FIN(E,N)",
  GEOGCS["ETRS89",
    DATUM["European Terrestrial Reference System 1989",
      SPHEROID["GRS 1980", 6378137, 298.257222101,
        AUTHORITY["EPSG", "7019"]],
      TOWGS84[0, 0, 0, 0, 0, 0, 0],
      AUTHORITY["EPSG", "6258"]],
    PRIMEM["Greenwich", 0,
      AUTHORITY["EPSG", "8901"]],
    UNIT["degree", 0.0174532925199433,
      AUTHORITY["EPSG", "9122"]],
    AUTHORITY["EPSG", "4258"]],
  PROJECTION["Transverse_Mercator"],
  PARAMETER["latitude_of_origin", 0],
  PARAMETER["central_meridian", 27],
  PARAMETER["scale_factor", 0.9996],
  PARAMETER["false_easting", 500000],
  PARAMETER["false_northing", 0],
  UNIT["metre", 1,
    AUTHORITY["EPSG", "9001"]],
  AXIS["Easting", EAST],
  AXIS["Northing", NORTH],
  AUTHORITY["EPSG", "3067"]]
gml_id: String (0.0)
```

```
nationalCode: Integer (0.0)
localId: String (10.0)
namespace: String (13.0)
nationalLevel: String (8.0)
LocalisedCharacterString: StringList (0.0)
Country: String (2.0)
language: StringList (0.0)
text: StringList (0.0)
script: StringList (0.0)
```

Listauksen tulkinta

- Tasolle kuuluu 362 aluekohdetta
- Koordinaattijärjestelmä on ETRS89-TM35FIN (EPSG:3067)
- Kohteilla on 10 eri ominaisuustietoa
- Neljä näistä on tyyppiä StringList
- nationalCode on tulkittu kokonaisluvuksi, vaikka kokenut paikkatietoihminen ehkä muistaa, että kuntakoodit ovat numeroista koostuvia merkkijonoja, joissa voi olla myös etunollia. Esimerkiksi Helsingin koodi on 091. Vilkaisu skeematiedostoon <http://xml.nls.fi/inspire/2010-04/AdministrativeUnits.xsd> varmistaa, että GDAL on tehnyt tässä kohdassa virheen.

StringList-tyyppiin ominaisuustietoihin kannattaa kiinnittää huomiota. Merkkijonolistamuoto merkitsee sitä, että esimerkiksi LocalisedCharacterString-ominaisuuden arvoksi voidaan antaa useita vaihtoehtoisia merkkijonoja. Kentän arvona voi olla esimerkiksi ”Maakunta,Landskap,Region” tai ”Kunta,Kommun,Municipality”.

Tavalliset paikkatieto-ohjelmistot eivät pysty käsittelemään merkkijonoluetteloita, mutta myöhemmin selviää, miten tämä ongelma voidaan kiertää.

Tietotyyppivirheen ”nationalCode – String” korjaaminen

Ei ole pahitteeksi ymmärtää, miksi GDAL tulkitsee, että nationalCode on tietotyyppiä Integer, vaikka se on .xsd -tiedostossa määritelty olevan String. Tämä johtuu siitä, että jos GDAL törmää .xsd -tiedostoa lukiessaan johonkin kohtaan, jota se ei pysty muuntamaan oman sisäisesti käyttämänsä tiedomallin mukaiseksi, se hylkää .xsd -tiedoston kokonaan. Sen sijaan GDAL käy systemaattisesti läpi varsinaisen XML-datatiedoston alusta loppuun ja etsii tiedostosta sellaiset elementit, jotka se tunnistaa, ja tekee niistä oman skeemansa. Tämä skeema tallentuu levyllä tiedostoksi, jonka tunnisteena on .gfs.

Maanmittauslaitoksen Inspire-määritysten mukaista XML-aineistoa tulkitessaan GDAL huomaa muun muassa, ettei se pysty tulkitsemaan, mitä sen pitäisi tehdä elementille Role – Enclave. Myös elementti ”inspireId”, jonka tyyppi on "base:IdentifierPropertyType" on sille käsittämätön. Viimeistään tässä vaiheessa GDAL hylkää .xsd -tiedoston kokonaan ja alkaa tutkia suoraan .xml -tiedostoa. Se löytää ominaisuustiedon nimeltään nationalCode, ja huomaa, että sen arvo on aina pelkkiä numeroita, ja tulkitsee tämän perusteella tietotyypin olevan kokonaisluku.

Jos haluaa, että nationalCode:n tietotyyppiä tulee muunnoksissa oikea, eli merkkijono, niin ogrinfo-komennon suorituksen yhteydessä syntyvä .gfs-tiedosto pitää avata tekstinkäsittelyohjelmaan ja korjata tämä kohta:

```
<PropertyDefn>
  <Name>nationalCode</Name>
  <ElementPath>nationalCode</ElementPath>
  <Type>Integer</Type>
</PropertyDefn>
```

tällaiseksi:

```
<PropertyDefn>
  <Name>nationalCode</Name>
  <ElementPath>nationalCode</ElementPath>
  <Type>String</Type>
</PropertyDefn>
```

Huomautus

Maanmittauslaitoksen XML-tiedostossa ainakin marraskuun 9. päivä päivitetystä versiossa kuntanumerot ovat ilman etunollia. MML ei ole tätä kirjoitettaessa vielä vastannut, onko kyseessä tietovirhe vai harkittu teko.

Spatialite-muunnoksen tekeminen ogr2ogr-ohjelmalla

XML-tiedoston muuntamiseen Spatialite-tietokannaksi tarvitaan lopulta vain yksi komento

```
ogr2ogr -f SQLite -splitlistfields -dscv spatialite=yes -dscv init_with_epsg=yes
mml_kuntajako_2011_10k.sqlite SuomenKuntajako_2011_10k.xml
```

Komennon parametrien selostus

- ogr2ogr on ohjelma, jolla muunnos tehdään
- -f SQLite kertoo, että tulos halutaan Spatialite-muotoon.
- -splitlistfields antaa ohjeen jakaa listamuotoiset ominaisuustiedot omiksi erillisiksi ominaisuustietokenttään, joita on helpompi käsitellä useimmilla tavallisilla paikkatieto-ohjelmistoilla. Spatialite-ohjain tukee myös StringList-tietotyyppiä siten, että esimerkiksi kuntajakoaineiston "Language"-kenttään kirjoitetaan arvoluettelo, kuten "2:fin,swe". Jos kenttä pilkotaan, niin syntyy tässä tapauksessa kaksi kenttää, joilla kummallakin on vain yksi arvo: "Language1"="fin" ja "Language2"="swe".
- Kaksi kohdetiedoston luomiseen vaikuttavaa parametria (-dscv, data store creation option) liittyvät Spatialite-tiedokannan luomiseen, eikä niiden merkitystä selitetä tässä.
- mml_kuntajako_2011_10k.sqlite on kohdetiedosto
- SuomenKuntajako_2011_10k.xml on lähtötiedosto

Shapefile-muunnoksen tekeminen ogr2ogr-ohjelmalla

Shapefile-muotoon muuntaminen tästä lähtöaineistosta ei ole yhtä suoraviivaista. Shapefile ei voi sisältää kuin yhden karttatason, joten muunnos pitää tehdä kahteen kertaan. Shapefile ei myöskään tue StringList-kenttiä, joten muunnoksessa on pakko käyttää parametria -splitlistfields, tai muussa tapauksessa näitä kenttiä ei muunneta ollenkaan. Shapefilemuunnoksen komennot ovat siis seuraavat:

```
ogr2ogr -splitlistfields -f "ESRI Shapefile" mml_kuntajako_2011_10k_alueet.shp
SuomenKuntajako_2011_10k.xml AdministrativeUnit
```

```
ogr2ogr -f "ESRI Shapefile" -splitlistfields mml_kuntajako_2011_10k_viivat.shp
SuomenKuntajako_2011_10k.xml AdministrativeBoundary
```

Huomautus

Jostain syystä yllä olevat komennot eivät toimineet, kun testasin niitä tätä ohjetta kirjoitettaessa. Parametrin -f "ESRI Shapefile" käyttäminen johti aina virheeseen. Tässä erikoistapauksessa virhe on helppo kiertää, koska shapefile on ogr2ogr-ohjelman oletustiedostomuoto. Seuraavat komennot toimivat testatessa.

```
ogr2ogr -splitlistfields mml_kuntajako_2011_10k_alueet.shp
SuomenKuntajako_2011_10k.xml AdministrativeUnit
```

```
ogr2ogr -splitlistfields mml_kuntajako_2011_10k_viivat.shp
SuomenKuntajako_2011_10k.xml AdministrativeBoundary
```

Koska shapefilen ominaisuustiedon nimessä voi olla korkeintaan 10 merkkiä, niin muunnos katkaisee muutamien kuntajakotiedostojen kentän nimiä, kuten LocalisedCharacterString. Alla oleva listaus näyttää shp-tiedostoon syntyvien ominaisuustietojen nimet ja tietotyypit.

```
gml_id: String (80.0)
nationalCo: String (10.0)
localId: String (10.0)
namespace: String (13.0)
nationalLe: String (8.0)
LocalisedC: String (36.0)
Localise_1: String (36.0)
Localise_2: String (36.0)
Country: String (2.0)
language1: String (3.0)
language2: String (3.0)
text1: String (52.0)
text2: String (52.0)
script1: String (4.0)
script2: String (4.0)
```

Muunnos Mapinfo-muotoon

Muunnos MapInfo tab-muotoon onnistuu seuraavilla komennoilla

```
ogr2ogr -f "MapInfo file" -splitlistfields mitest.tab
SuomenKuntajako_2011_4500k.xml AdministrativeUnit -nln AdministrativeUnit
```

```
ogr2ogr -f "MapInfo file" -splitlistfields mitest.tab
SuomenKuntajako_2011_4500k.xml AdministrativeBoundary -nln
AdministrativeBoundary
```

Huomautuksia GDAL:in MapInfo-ajurista ja käytetyistä parametreistä

- MapInfo ei tue StringList-tietotyyppiä, joten -splitlistfields -parametrin käyttäminen on pakollista.
- Ogr2ogr ei voi kirjoittaa olemassa olevaan tab-tiedostoon. Siitä syystä XML-tiedostoon sisältyvät kaksi tasoa on muunnettava omiksi tab-tiedostoikseen (kun ensimmäinen taso on kirjoitettu, tab on olemassa, eikä toisen tason lisääminen siihen onnistu).
- -nln (new layer name) parametrilla voidaan määrätä, mikä niminen karttataso tab-tiedostoon syntyy. Oletus on sama kuin tab-tiedoston nimi.