

sorvi työkaluja avoimelle datalle

Leo Lahti* ja Juuso Parkkinen†

26. lokakuuta 2011

1 Johdanto

Vapaasti saatavilla olevien tietoaaineistojen määrä on ollut voimakkaassa kasvussa [2]. Talouteen, säähän, liikenteeseen, koulutukseen ja muihin alueisiin liittyvää dataa on alettu avaamaan julkishallinnon toimesta Suomessa ja muualla. Aineistojen pöyhintä voi avata uusia näkökulmia ja avoimuus mahdollistaa yhdistämisen toisiin tietolähteisiin, jolloin voidaan vastata kysymyksiin joihin yksittäisillä aineistoilla ei päästä käsiksi. Siten laaja saatavuus ja käyttö voi merkittävästi nostaa datan arvoa.

Laskennallisten työkalujen saatavuus on osoittautunut pullonkaulaksi avointen datojen laajemmalle käytölle. Tämä R-paketti pyrkii paikkaamaan puuttava tarjoamalla yleiskäyttöisiä välineitä avointen rajapintojen kautta saatavilla olevien julkisten tietoaaineistojen automatisoituun lataamiseen, siistimiseen, yhdistelyyn, louhintaan ja visualisointiin. Pakettiin pyritään kokoamaan kattava kokoelma eri lähteistä koottuja ratkaisuja avoimen datan käsittelyyn. Uusia ehdotuksia ja lisäyksiä otetaan mieluummin vastaan.

Yksityiskohtaisempia esimerkkejä paketin käytöstä Suomi-datan penkomiseen löydät Louhos-blogista¹.

2 Paketin asennus

Paketin saa asennetuksi suoraan R:stä käsin. Paketin riippuvuudet saattavat kuitenkin ensin tuottaa lisäsäätöä. Tässä asennusesimerkit ja kommentteja näidenkin osalta, toimii ainakin Ubuntussa. XML-paketti vaatii 'libxml2-dev'-palikan asennusta esim pakettienhallinnasta; rgdal- ja rgeos-pakettien asennus edellyttää GEOS:n asennusta (<http://trac.osgeo.org/geos>) esim. asentamalla pakettienhallinnasta 'geos-dev', 'libgeos-dev' riippuvuuksineen ja manuaalisesti PROJ.4

*Helsingin yliopisto <leo.lahti@iki.fi>

†Aalto-yliopisto <juuso.parkkinen@gmail.com>

¹<http://louhos.wordpress.com>

(<http://trac.osgeo.org/proj>); rgdal-paketti vaatii 'libgdal-dev'-palikkaa. Nyt riippuvuuksien ja esimerkeissä käytettyjen pakettien pitäisi asentua R:ssä komennolla:

```
> install.packages(c("methods", "ggplot2", "gpclib", "mapproj",  
+ "maps", "maptools", "pxR", "RCurl", "ReadImages", "RgoogleMaps",  
+ "rjson", "sp", "XML"))
```

Muita hyödyllisiä paketteja, joita osa esimerkeistä käyttää ja joita kannattaa asentaa tarpeen mukaan ovat

```
> install.packages(c("gdata", "plyr", "png", "rgdal", "rgeos"))
```

Nyt sorvin pitäisi asentua R:ssä:

```
> install.packages("sorvi", repos = "http://R-Forge.R-project.org")
```

3 Pakettiin viittaaminen

Paketin tarjoamat välineet ovat vapaasti käytettävissä FreeBSD-lisenssillä². Mikäli paketista on apua, toivomme viittaamista työhön [1]. Lisää tietoa löytyy projektin kotisivulta³.

4 Julkiset tietokannat

Alla esimerkkejä joidenkin julkisten aineistojen lataamisesta sorviin.

4.1 Maanmittauslaitos

Suomen karttatietoja on haettu Maanmittauslaitoksen sivuilta⁴ rajapintasyistä valmiiksi sorviin. Maanmittauslaitoksen aineistoja⁵ voi selata system.file-funktion avulla:

```
> dir(system.file("extdata/Maanmittauslaitos/", package = "sorvi"))
```

Shape-muotoisen aineiston voi ladata komennolla

```
> file.name <- "extdata/Maanmittauslaitos/1-milj-Shape-etrs-shape/forest.shp"  
> shape.file <- system.file(file.name, package = "sorvi")  
> sp <- readShapePoly(shape.file)  
> sp <- putsaa.shape.mml(sp)
```

²http://en.wikipedia.org/wiki/BSD_licenses

³<http://sorvi.r-forge.r-project.org/>

⁴<http://www.maanmittauslaitos.fi/aineistot-palvelut/digitaaliset-tuotteet/ilmaiset-aineistot/hankinta>

⁵Jatkokaytto sallittu - (C) MML 2011 - <http://www.maanmittauslaitos.fi/node/6417>

4.2 Tilastokeskus

Tilastokeskuksen sivuilta⁶ löytyy avoimia aineistoja PC-Axis muodossa. Tässä esimerkkinä poimittu kuntien mediaanitulot (määrittele muuttujaan url osoite "http://pxweb2.stat.fi/Database/StatFin/tul/tvt/2009/120_tvt_2009_2011-02-18_tau_112_fi.px"):

```
> library(pxR)
> px <- as.data.frame(read.px(url))
> px <- putsaa.px(px)
```

4.3 Muuta

4.3.1 Postinumerot

```
> suomen.postinumerot <- hae.postinumerot()
```

4.3.2 Vaestötiheys maakuntatasolla

Poimi väestötiheystiedot Wikipediasta:

```
> url <- "http://fi.wikipedia.org/wiki/V%C3%A4est%C3%B6tiheys"
> vaestötiheys <- hae.maakuntatiedot(url)
```

5 Versiotiedot

```
> sessionInfo()
```

```
R version 2.13.0 (2011-04-13)
Platform: i686-pc-linux-gnu (32-bit)
```

```
locale:
```

```
[1] LC_CTYPE=fi_FI.utf8      LC_NUMERIC=C
[3] LC_TIME=fi_FI.utf8      LC_COLLATE=fi_FI.utf8
[5] LC_MONETARY=C           LC_MESSAGES=fi_FI.utf8
[7] LC_PAPER=fi_FI.utf8     LC_NAME=C
[9] LC_ADDRESS=C            LC_TELEPHONE=C
[11] LC_MEASUREMENT=fi_FI.utf8 LC_IDENTIFICATION=C
```

```
attached base packages:
```

```
[1] stats      graphics  grDevices  utils      datasets  methods   base
```

Viitteet

- [1] Leo Lahti ja Juuso Parkkinen (2011). sorvi - avoimen datan työkalupaketti
URL: <http://louhos.wordpress.com>

⁶<http://www.stat.fi/tup/tilastotietokannat/index.html>

- [2] Antti Poikola, Petri Kola ja Kari A. Hintikka (2010). Julkinen data – johdatus tietovarantojen avaamiseen Liikenne- ja Viestintäministeriö. Edita Prima Oy, Helsinki 2010. URL: <http://www.julkinendata.fi>