Linked Data as a model for increased data interoperability in Finland

In general, there is a problem in integrating of statistics and geospatial data, because the link between them is missing and the current production lines of statistics and geospatial data are separated. As a solution to this, the areal classifications used in area statistics production could be used systematically as links between the area statistics and corresponding geographies.

Another problem is that because geospatial data and classifications are managed by different organisations, there are partly overlapped data productions. This solution provides a great opportunity to combine data productions in two different organisations to one seamless unified process.

Description of the problem

There is a problem in combining statistical areal classifications and corresponding geographical data in a systematic way at Statistics Finland (hereafter STAT-FI). The classification correspondence tables are utilised, when producing municipality-based statistical units, but the geographical data and classifications are not systematically linked.

Another problem is the duplicated work between two organisations that are using the same data, but different versions. The National Land Survey of Finland (hereafter NLS) is producing municipalities as geographic data objects (polygons, lines) in multiple resolutions. For the needs of STAT-FI, the NLS has to do a specific customised product of municipalities. STAT-FI still needs to customise the data further to produce municipality-based statistical units.

The solution to both of these problems is a completely new idea of a production model between the organisations that utilises linked data and unique identifiers. In the model the area classifications and geographies are combined as linked data through unique identifiers and they are served from both organisations' own services in machine readable format. Since a link is created between the data, the data can be queried in the format that is suitable for each use case. Also, the data would be available to customers.

Solution

The solution to the problem may be as shown in the picture below:

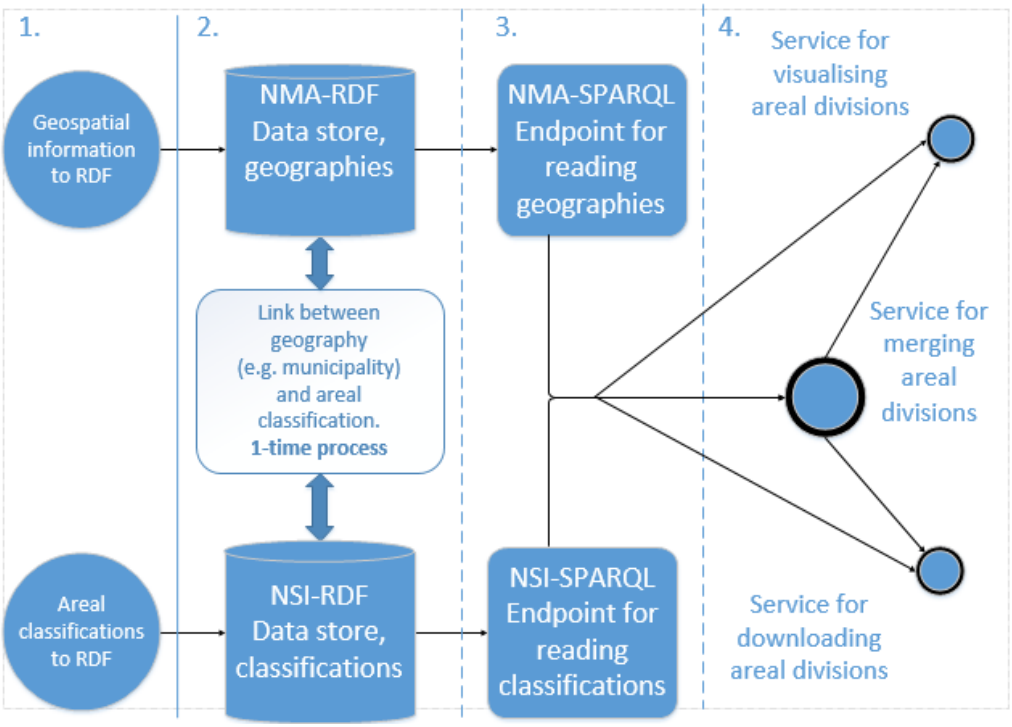

Figure 1: Suggested production model

1. In the picture, both organisations would have their data stores for RDF data. First, the NLS needs to transform their geospatial data to RDF and download it to the NMA-RDF Data store, (National Mapping Agency RDF Data store). STAT-FI would also have to transform areal classifications to RDF format and download it to the NSI-RDF Data store (National Statistical Institute RDF Data store).

2. In the second phase, the linkage between the different data sources is created. The links are described in the used RDF ontologies. To secure two-way linking, they can be included in both ontologies. When new municipality data or areal classification are provided in the RDF data stores, the links will be created. For creating the linkage, the unique identifiers in both data would be utilised.

3. The third phase is to disseminate the data through SPARQL endpoints. Both organizations have their own endpoints and data could be queried from either SPARQL endpoint using federated query.

4. The data can be used directly from the SPARQL endpoint or a custom-made service could be created on the top of the endpoint. The service can be used for data visualisation, for processing the data to analysis purposes or to transform the data in different data formats.

Result

Solutions in production use are not yet available. But the production model was used in a proof-of-concept, POC. In the POC both the classifications and the geographies were transformed into RDF and links were created between them utilising the unique identifiers of the data. Classifications and geographies were in two separate RDF stores, with two separate SPARQL endpoints. To utilise the data, a new service called ALLUsion was created. In the service it was possible to merge geometries of the areas, query statistical data of the areas and have the results visualised on a map and the statistics in a table.

Below are some examples of querying combinations of geographic data and areal class.

Figure 2: Geographic data and information box of the Helsinki-Uusimaa major region.

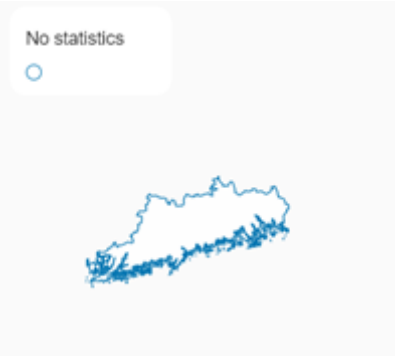

Figure 3: Helsinki-Uusimaa major region in more detailed geometry.

Below is an example of querying combination of geographic data, areal class and statistical data.

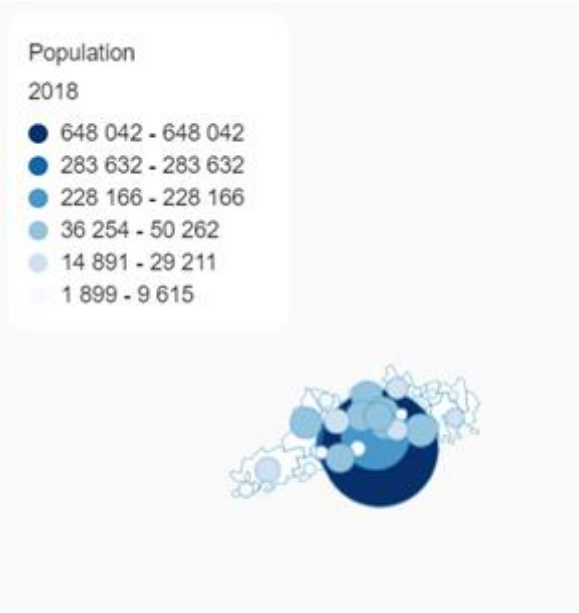

Figure 4: Population 2018 in the municipalities of Uusimaa region.

More information

The source code of ALLUsion service is availabe in GitHub:
https://github.com/StatisticsFinland/allusion

Contact information

Tuuli Pihlajamaa, Statistics Finland, tuuli.pihlajamaa@stat.fi or sijaintipalvelut@stat.fi

Eero Hietanen, National Land Survey of Finland, eero.hietanen@nls.fi