Antti Santaharju

15 May 2020

Statistics Finland's geospatial data warehouse

1.GSGF principles and Statistics Finland's geospatial data warehouse

Statistics Finland's geospatial data warehouse is one example of how to integrate statistical and geospatial data. Figure below shows how this solution implements the GSGF principles.

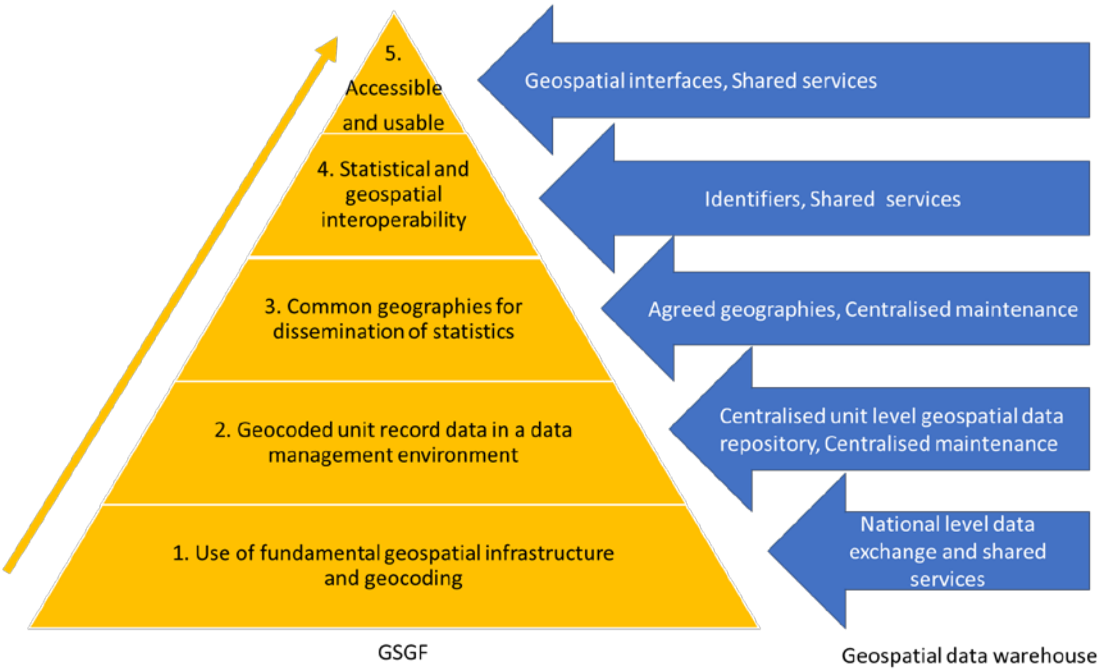

5. Accessible and usable
4. Statistical and geospatial interoperability
3. Common geographies for dissemination of statistics
2. Geocoded unit record data in a data management environment
1. Use of fundamental geospatial infrastructure and geocoding
GSGF
Geospatial interfaces, Shared services
Identifiers, Shared services
Agreed geographies, Centralised maintenance
Centralised unit level geospatial data repository, Centralised maintenance
National level data exchange and shared services
Geospatial data warehouse

Figure1: The principles of the Global Statistical Geospatial Framework (GSGF) realized in Statistics Finland's geospatial data warehouse

2. What has been the motivation to execute the method? (advantages, benefits)

Geospatial information has been identified as one logical data repository in Statistics Finland's Enterprise Architecture (Figure 2). In order to understand its content and connections to other data repositories a geospatial reference architecture has been designed. It describes the processing of the geospatial information utilised and produced in Statistics Finland's statistical production.

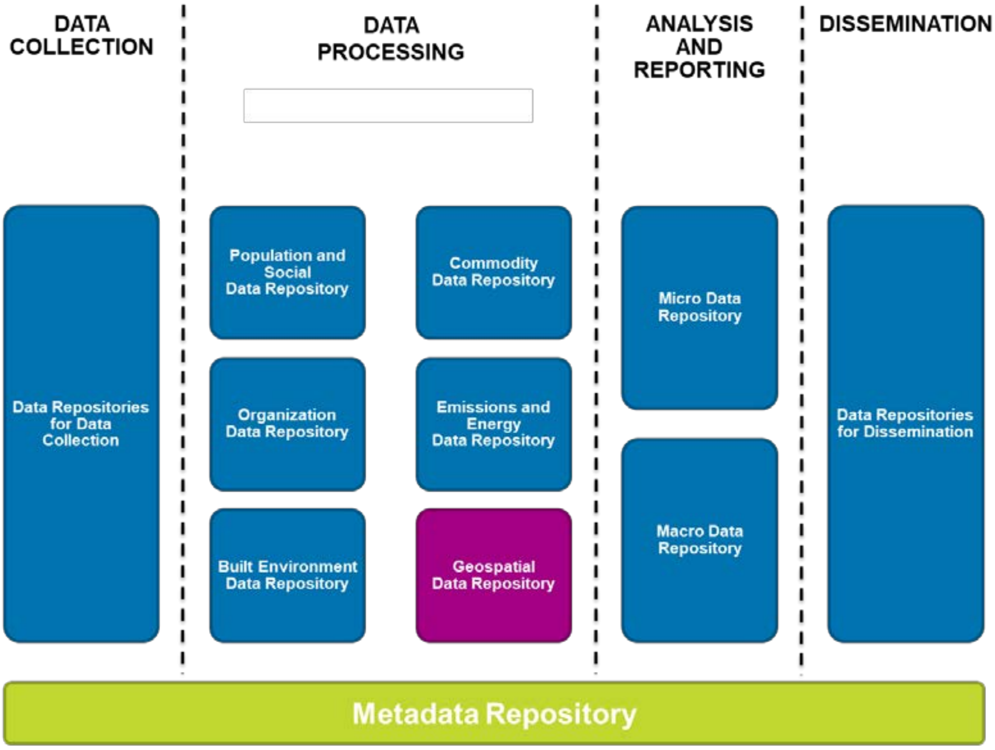

DATA COLLECTION
DATA PROCESSING
ANALYSIS AND REPORTING
DISSEMINATION
Data Repositories for Data Collection
Population and Social Data Repository
Commodity Data Repository
Organization Data Repository
Emissions and Energy Data Repository
Built Environment Data Repository
Geospatial Data Repository
Micro Data Repository
Macro Data Repository
Data Repositories for Dissemination
Metadata Repository

Figure 2. Statistics Finland's information architecture

Previously, the processing of geospatial information has been implemented with overlapping statistics-specific solutions that maay have led to mutually non-uniform results and overlapping work. Decentralised data warehousing has allowed the same data being stored in several locations (databases or separate files on drives). Also, the tools and methods used may have been statistics-specific stand-alone solutions. From this background, it has been difficult to utilise the national SDI by standardised solutions. It has also complicated defining Statistics Finland's own role and responsibilities as a member of the national SDI.

Statistics Finland's geospatial reference architecture is the answer to the challenges described above. Its key perspective is to improve interoperability within Statistics Finland and outside the agency in the operating environment of geospatial information. The reference architecture shows how the activity related to geospatial information should be arranged in order that:

- Statistics Finland attains a centralised operating model for the production of geospatial information and processing of geospatial information in the other statistical production processes
- Geospatial information is defined in the logical geospatial data warehouse
- Statistics Finland cooperates with other producers of geospatial information. Geospatial information from other producers is used whenever possible in accordance with the national division of responsibilities.

The implementation work has been started with the following results so far:

- Geospatial data warehouse was introduced in June 2019. It includes CSPA compatible services for the processing and integration of geospatial information.

- A management group for geospatial information has been established.
- Statistics Finland launched closer cooperation with the National Land Survey of Finland to achieve a compatible enterprise architecture between the organisations. Cooperation enables the use of services provided by the National Land Survey of Finland in the statistical production process, e.g. for geocoding and coordinate conversions.

3. How can the data integration method be described?

According to the Statistics Finland’s geospatial reference architecture, all Statistics Finland’s geospatial information is produced only once and stored in the geospatial data repository. Geospatial information from other producers is used whenever possible. *NMA Finland* produces municipal based geometries and *Digital and Population Data Services Agency* produces address and coordinate information for all Finnish buildings.

All geospatial information required for statistical production is available from Statistics Finland’s Geospatial Data repository. Actors of statistical production processes (e.g. statistical production processes) link their statistical units once to the objects of the geospatial data warehouse and store the geospatial object Id in their databases. After statistical units are linked to the geospatial objects[1], repetitive processes can extract up-to-date geospatial information from this repository (Figure 3). E. g. geometries are updated in the Geospatial data repository whenever MML updates them. Also, the building area links are updated according to these geometries. Object Id can be used as linking variable to integrate data from different phenomena (in different databases) by location.

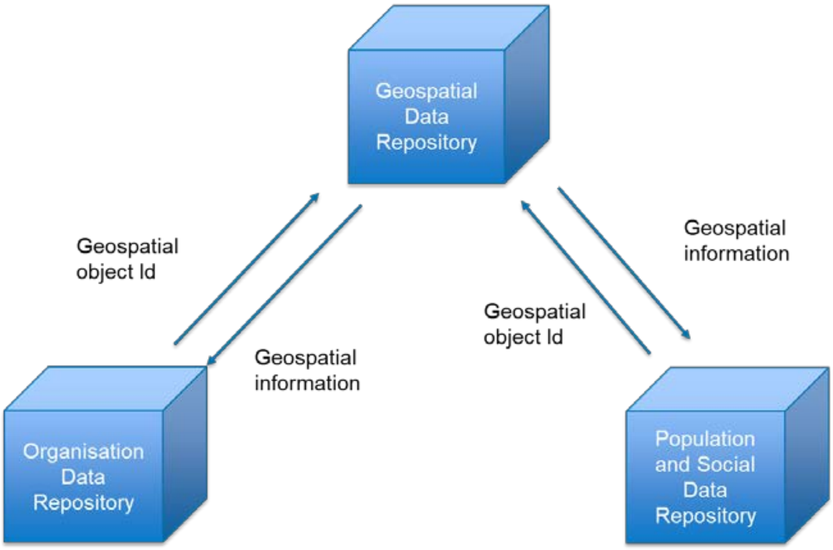


Figure 3 geospatial data repository and its connections to the other data repositories.

The content of the geospatial data repository covers

- All geometries used in statistical production

[1] Geospatial data repository provides tools for linking statistical units to the geospatial objects. Linking process is explained below.

- All Finnish Buildings and their Id's, addresses, coordinates, building-location links (derived from geometries)

The content of the geospatial data repository is described in more detail in Annex 1.

Geospatial data repository provides tools (shared services) for linking statistical units to the geospatial objects and tools for processing geospatial data. Examples of these services are shown in table 1.

Service	Input (example)	Output (example)
Address standardization	Address as a string ("address": "Työpajankatu 13 F 2 00580 Helsinki)	Standardized address in components ("streetName": "Työpajankatu", "streetNumber": "13", "streetLetter": "F", "apartmentNumber": "2", "Postal code": "00580", "Municipality": "Helsinki")
Address editing	Misspelled address ("streetName": "Työjankatu", "streetNumber": "13", "streetLetter": "F ", "postalCode": "00580")	Address corrected by fuzzy logic. Properly written address and confidence of the output ("streetName": "Työpajankatu", streetNumber": "13", "streetLetter": "F", "postalCode": "00580", "confidence": 88)
Geocoding/address linking	Address	Coordinates, Object Id
Building service	Object Id	Address and area information
Distance calculation	Two locations expressed in coordinates	Distance between locations (straight distance by road)
Coordinate conversions	Coordinates in a given coordinate system	Coordinates in another coordinate system

Table 1. Statistics Finland's shared services for geospatial processing

Actors of statistical production processes link their statistical units to the objects of the geospatial data warehouse. Automated linking process shown in figure 4 is based on the services documented above. Linking process has the following steps.

1. Addresses must be standardised in the national standard format in which they are stored in the geospatial data repository. This is performed using the address standardization service which divides the address into standard compliant components.
2. When the addresses are in the common standardized format, they can be linked to the addresses of the geospatial data repository. This is performed using the geocoding/address linking service. Most of the

addresses will be linked directly. For them, the process is completed. The Process receives a geospatial object Id’s for the linked addresses in the response.

3. Some share of addresses processed in step 2 do not link. Non-linking is due the typos or other differences in spelling. Process performs address editing service for these addresses. It uses fuzzy logic to link misspelled addresses to the correct ones in Geospatial data repository. Service proposes the most probable valid address for the misspelled ones and confidentiality rate of this proposal.
4. The output of the process step 3 are the edited addresses. Geocoding/address linking service is performed again for this output. Allmost all the addresses will be linked after they are edited, and the process is completed for these addresses. The Process receives a geospatial object Id’s for the linked addresses in the response.
5. If some addresses are unlinked after the process step 4, they can be stored in the Geospatial data repository by Missing address update -service. These addresses are flagged as invalid. Invalid addresses are stored in the geospatial data repository because it should be the only source for geospatial information.

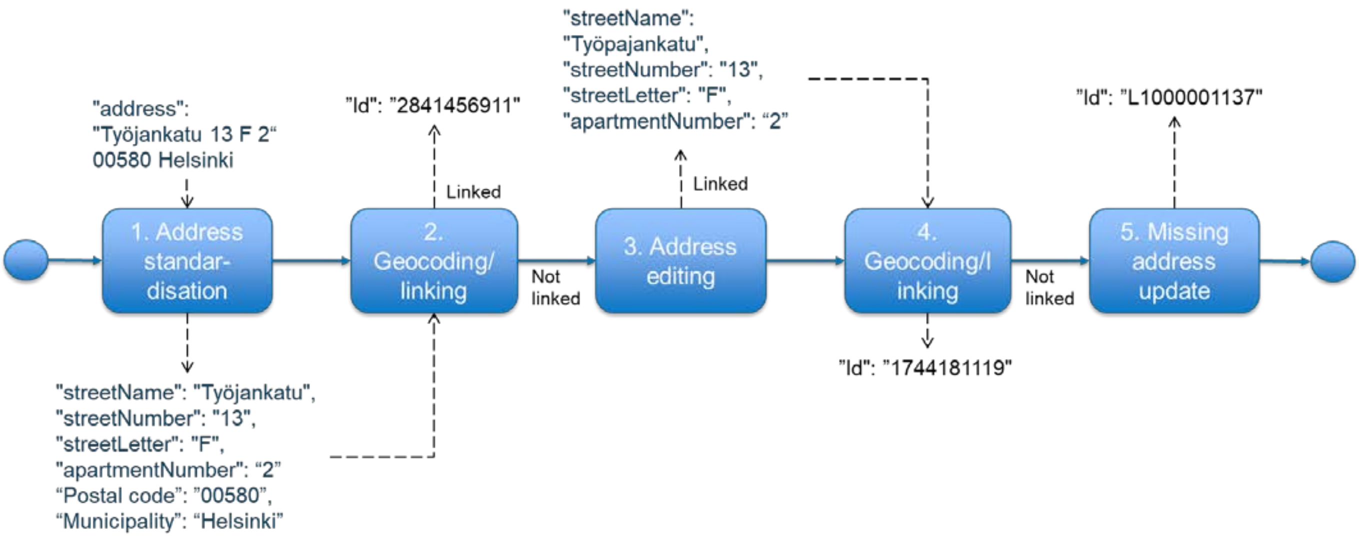


Figure 4 Automated linking process for geospatial objects

Once the addresses are linked into the geospatial data repository the statistical processes of different statistical domains are able to

- Extract up-to-date geospatial information from the geospatial data repository which is maintained by geospatial management team
- Fully utilise the shared services for the processing of the geospatial data
- Use geospatial object Id’s for the integration of their data to the data from other domains by location

4. The issues identified that should be tackled

The key challenge is that existing statistical production processes need to change when they adopt the centralised operating model. Although the benefits are clear, changes require work. Attempts are made to support this by providing tools and best practice cases as benchmark. Users should be involved in the planning from the beginning. This can be organised through regular review meetings. Geospatial management team has also set up a virtual discussion channel for related discussions. It is worth noting that statistics are losing some amount of flexibility when they adopt centralised solutions.

Statistics Finland’s geospatial data warehouse has not been fully implemented in the statistics Finland’s statistical production processes. First adoptions have been made and implementation plans have been defined for the key processes of the social and organisational data repositories.

Centralised operating model requires that maintenance and development issues are well organised. Statistics Finland has established a Geospatial Management Group. This Group and its tasks are described in the following chapter. Statistics Finland’s experience is that technical change management of the shared geospatial services must be well resourced and organised. This needs to be further developed so that the issues presented in this document work well in Statistics Finland's processes.

5. Recommendations for the management level from the example

Centralised operating model presented in chapter 3 requires that also the maintenance and management of geospatial data and geospatial issues are centralised and well organised. Statistics Finland has established a Geospatial Management Group. This can be used as an example how the centralised operating model can be established.

Tasks of the statistics Finland’s geospatial management group are:

- Contact point (statistical production, customers)
- Maintenance of the shared geospatial data warehouse
- Maintenance of shared geospatial services
- Owner of the geospatial reference architecture
- Geospatial related System and Technology Architecture
- Geospatial Methods
- Co-operation in the Geospatial field

Geospatial Management Group and its operating environment is shown in figure 6. The Members of the management group are highlighted in green colour. Other important internal stakeholders of the development, management and maintenance of the geospatial issues are highlighted in yellow colour. Geospatial Management Group provides services (listed above) for statistical production and statistical actors. Geospatial Management Group has its own steering group consisting of the directors of the IT-, Statistical Methods- and Social Statistics departments. Statistics Finland's external stakeholders are also shown in the figure. These stakeholders are highlighted in red colour.

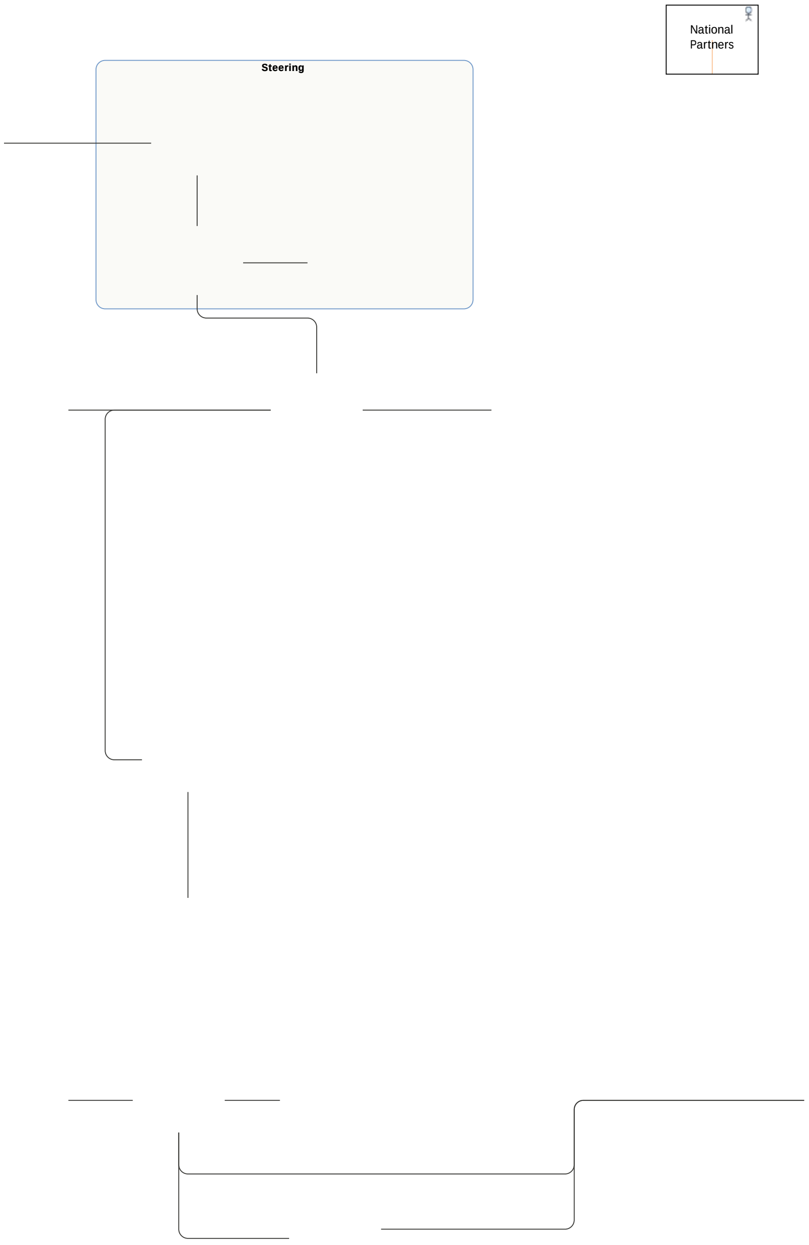


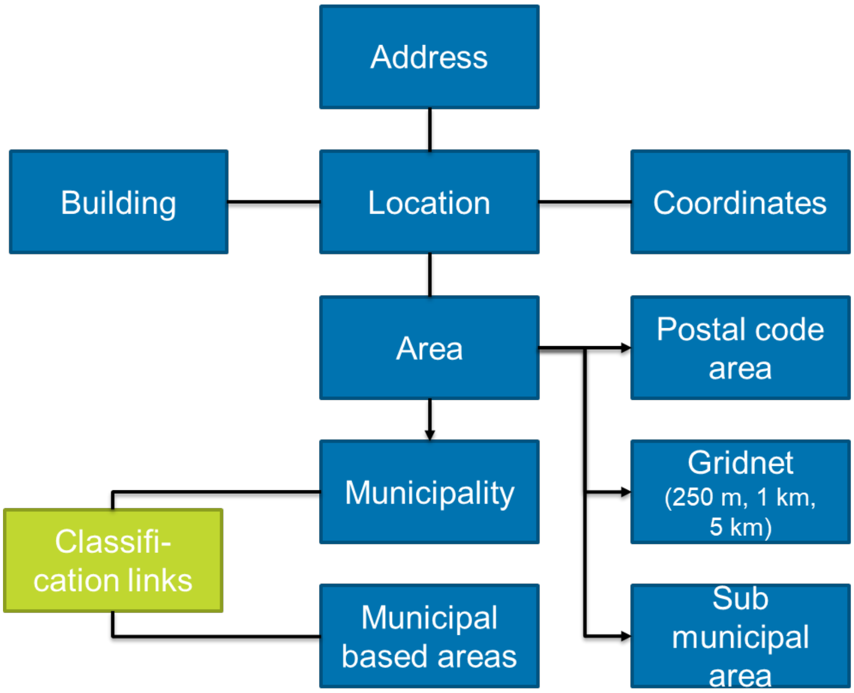
Figure 7. Actors of Statistics Finland’s Geospatial Management Group and its operating environment

6. Contact information

Antti Santaharju
Statistics Finland,
antti.santaharju@stat.fi

Rina Tammisto
Statistics Finland,
rina.tammisto@stat.fi

ANNEX 1 The content of the Statistics Finland’s geospatial data repository


Address
Building
Location
Coordinates
Area
Postal code area
Municipality
Gridnet (250 m, 1 km, 5 km)
Classifi-cation links
Municipal based areas
Sub municipal area