



GSGF Europe - Implementation guide for the Global Statistical Geospatial Framework in Europe

Annex 2: Good Practice Cases

Version 1.0

28 February 2019

Title: GSGF Europe - Implementation guide for the Global Statistical Geospatial Framework in Europe - Annex 2:
Good Practice Cases

Project: Eurostat ESSnet grant project GEOSTAT 3

Grant agreement number: 08143.2016.002-2016.752

It is permitted to copy and reproduce the content in this report. When quoting, please state the source.

© EFGS and Eurostat 2019

Content

1	Principle 1: Use of fundamental geospatial infrastructure and geocoding of statistical information.....	5
	(C1.1) Dutch Base Registry for Addresses and Buildings (Netherlands).....	5
	(C1.2) Geocoded address access points – data collection & provision (Austria)	8
	(C1.3) Building Points and the National Dwellings Register (Portugal).....	12
	(C1.4) A geocoded building and dwelling register as a base for the production of geospatial statistics (Switzerland).....	17
	(C1.5) The benefits of open address data (Denmark)	19
	(C1.6) The Real Property Register – authoritative location data for geocoding within the NSDI (Sweden).....	20
2	Principle 2: Geocoded unit record data in a data management environment	25
	(C2.1) Geocoding guidance prepared by ABS (Australia)	25
	(C2.2) Dutch Geocoding service API (Netherlands).....	26
	(C2.3) National Geocoding and Routing Services (Finland).....	28
	(C2.4) Geocoding Quality Declaration at Object Level (Norway).....	30
	(C2.5) Point-of-entry validation in data collection (Estonia).....	36
3	Principle 3: Common geographies for production and dissemination of statistics	40
	(C3.1) REGINA – Management of coding system for administrative geographies (Sweden)	40
	(C3.2) Open geography portal by ONS (United Kingdom).....	43
	(C3.3) European Location Services (ELS) – central access to harmonised, pan-European, authoritative geospatial information and services (EuroGeographics)	44
	(C 3.4) Exploring OGC Discrete Global Grid Systems DGGS (EFGS)	47
4	Principle 4: Statistical and geospatial interoperability – Data, Standards and Processes	50
	(C 4.1) Geospatial Reference Architecture – a tool for managed utilisation of geospatial information (Finland)	50
	(C 4.2) Making SDMX fit for INSPIRE – How statistical tools can deliver INSPIRE compliant data and metadata (Eurostat)	58
	(C 4.3) Publish data once and leave data at its source (Netherlands)	64
	(C 4.4) Linked Data based model of joint production process of statistical units (Finland)	68
	(C 4.5) Development of guidelines for publishing statistical data as linked open data (Poland)	72
5	Principle 5: Accessible and usable geospatially enabled statistics.....	77
	(C 5.1) PX-Web API adapter for the Oskari platform (Finland)	77

This annex is a part of the final report from the GEOSTAT 3 project. The use cases presented in this annex are linked to Chapter 2 of the main report and represent national good practices to illustrate actions underpinning the principles, requirements and recommendations of the Global Statistical Geospatial Framework and the GEOSTAT 3 implementation guide.

1 Principle 1: Use of fundamental geospatial infrastructure and geocoding of statistical information



(C1.1) Dutch Base Registry for Addresses and Buildings (Netherlands)

Keywords: Principle 1, principle 3, principle 4, address register, geospatial infrastructure, point-based foundation, persistent identifiers, open address data, framework

This use case refers to a number of requirements provided within different principles. Most significantly, it refers to the requirements and recommendations within Principle 1, to use point-based geospatial information from National Spatial Data Infrastructures for geocoding, to use unique identifiers and to identify roles and responsibilities of different data producing agencies and build institutional agreements and legislations (requirement 1.1 and 1.2). It also illustrates requirements provided within Principle 3, to maintain a consistent framework of national statistical and administrative geographies and Principle 4, to enable data integration through consistent semantics and concepts across domains and to explore the potential of Linked Open Data for increased interoperability.

Introduction

The Dutch Base Register for Addresses and Buildings (BAG) is part of the system of key registers in the Netherlands. The BAG is maintained by the municipalities and distributed centrally by the Dutch Cadaster and is one of the base registers at the moment.

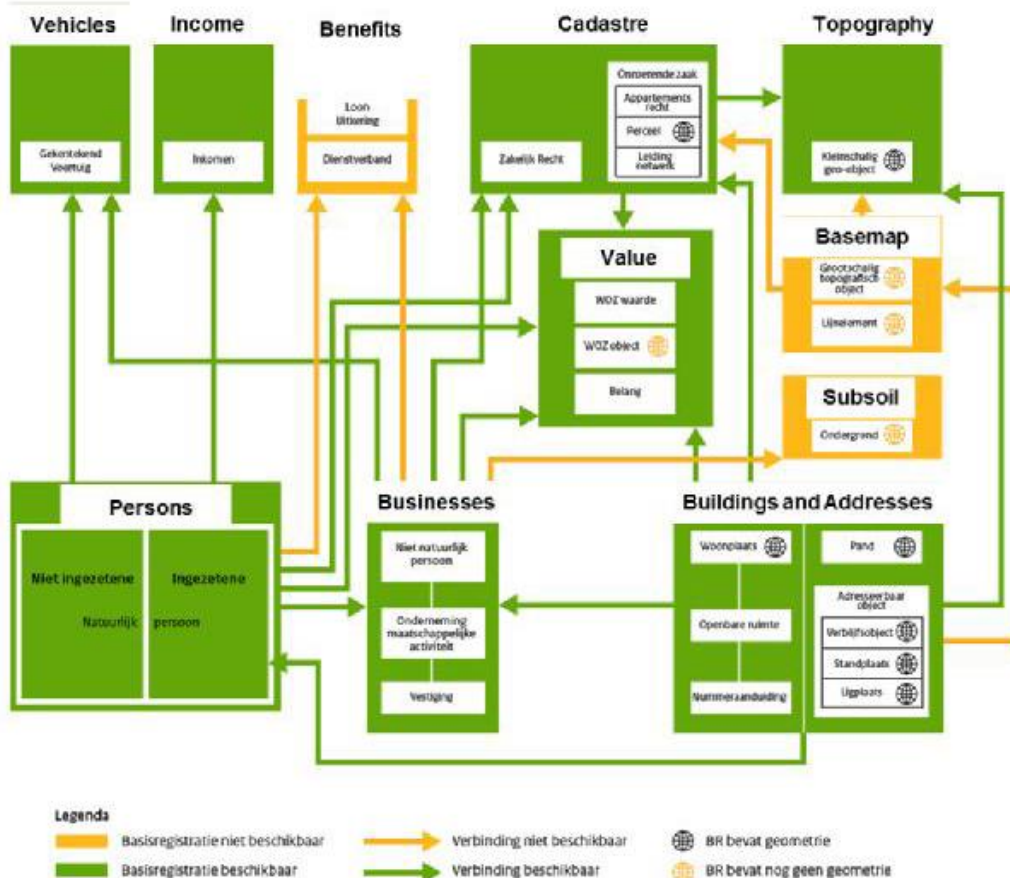


Figure 1: The Dutch Base Register for Addresses and Buildings (BAG)

Registers with globes contain geometry

Explanation of collars: green=ready and yellow= not yet ready

It is an authentic registration, which means that public authorities are obliged to use it, to prevent unnecessary extra work in maintaining address databases, copies and data versions. The information is open data. This is an effective implementation of the principle “storing data only once and leaving data at its source”.

BAG is updated daily and contains address information and building information including the geometries (address points, and building contours). The address points with their link to the population registry by entity ID are the basis of the spatial statistics of the Netherlands.

Base register of Addresses and Buildings (BAG)

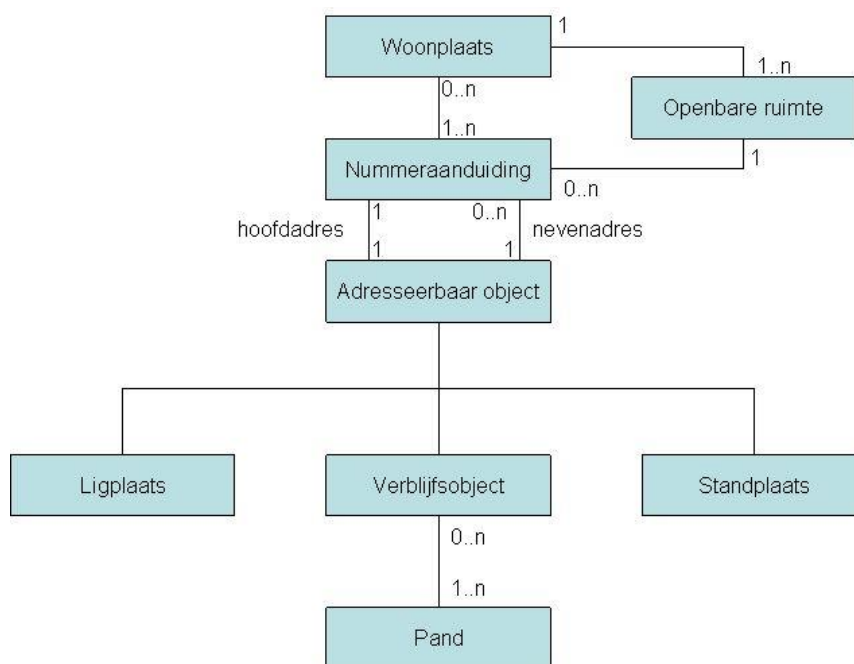


Figure 2: The structure of the BAG

<http://bag.kadaster.nl/def#OpenbareRuimte>

Public space: Street or Square

<http://bag.kadaster.nl/def#Brondocument>

Reference to the official document describing the change (not in this image)

<http://bag.kadaster.nl/def#Ligplaats>

Area reserved for houseboat

<http://bag.kadaster.nl/def#Standplaats>

Area reserved for caravan or trailer

<http://bag.kadaster.nl/def#Verblijfsobject>

Entity, smallest area where people live or work

<http://bag.kadaster.nl/def#Pand>

Building containing none, one or more entities

<http://bag.kadaster.nl/def#Nummeraanduiding>

Address with identifier.

<http://bag.kadaster.nl/def#Woonplaats>

City, town

Buildings contain one or more entities where people live or work. Buildings adjacent and part of the cadastral parcel may contain no entities.

Dwellings and work places are registered to an entity, an area for houseboat or an area for trailer, these are called “Adresseerbaar object”. The entity, houseboat and trailer will have at least one address (hoofdadres) but multiple addresses may occur (nevenadres). An address is only an attribute of an entity, houseboat or trailer.

All occurrences of the objects in the figure of the structure of the BAG have a unique identifier. Once added to the BAG, these identifiers will not change over time. The content may change, for instance, an address may change but the address identifier will not change and stay connected to the entity identifier.

Stelselcatalogus (all base registers)

The base registry of addresses and buildings is one of the base registries interconnected.

The content and description, definition of different elements of all base registries is in the “Stelselcatalogus” <https://www.stelselvanbasisregistraties.nl/registraties/>

The description and published elements for instance of “verblijfsobject” (smallest entity of working and living) is revealed through:

<https://www.stelselvanbasisregistraties.nl/begrippen/Adres/BAG/Verblijfsobject>

Linked data BAG

All BAG objects are served as Linked Data. The table above shows their translation in English, but also the link to their description.

The data is also served as OGC geo services: <https://bag.basisregistraties.overheid.nl/geo-services?articleid=1927964#8074e10f7b8333d716001c1b3a7348a3>

The address part and the buildings are served as separate INSPIRE services according to their corresponding INSPIRE themes:

[http://nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/search?facet.q=type%2Fdataset&isChild='false'&resultType=details&any_OR_title=adressen%20\(inspire%20geharmoniseerd\)&fast=index&content_type=json&from=1&to=20&sortBy=relevance](http://nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/search?facet.q=type%2Fdataset&isChild='false'&resultType=details&any_OR_title=adressen%20(inspire%20geharmoniseerd)&fast=index&content_type=json&from=1&to=20&sortBy=relevance)

This is a good example of the GSGF principle of following the standards and making data machine-readable.

Contact information

Drs. Niek van Leeuwen, Statistics Netherlands, n.vanleeuwen@cbs.nl or info@cbs.nl

Ir. Pieter Bresters, Statistics Netherlands, p.bresters@cbs.nl or info@cbs.nl

(C1.2) Geocoded address access points – data collection & provision (Austria)

Keywords: Principle 1, address register, building register, geospatial infrastructure, point-based foundation, routing

This use case refers to all of the recommendations provided within Principle 1, to use fundamental geospatial infrastructure and geocoding of statistical information. In addition, some recommendations concerning Principle 2, Geocoded unit record data in a data management environment are also addressed.

Introduction

The use of geocoded parcel addresses, as well as buildings, showed the need for precise access points. For purposes such as routing and navigation, it is essential that the “last mile” is routed correctly. Therefore, with the launch of the Austrian Graph Integration Platform GIP (the authoritative road network), the idea of meaningful address access points arose, to create operational coordinates for routing purposes. The solution was to move the address coordinate to a position on the driveway, close to the entrance, as address access point and deriving a new coordinate from that directly on the graph of the connecting road section (GIP coordinate).

Description of problem


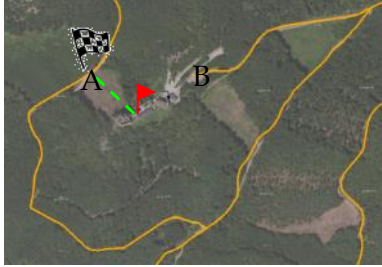

With the start of the Address-, Buildings and Dwellings Register (A-BDR) in 2004, there were two types of coordinates: the “address coordinates” (referring to parcel addresses) and “building coordinates” (referring to building addresses). Originally, address coordinates were mostly used to identify the addresses and were automatically placed at the label points of the parcel. Equally building coordinates originally were placed anywhere inside the buildings or in some cases they were even identical to the corresponding address coordinates. This was a good start to get an almost complete set of geocoded addresses and buildings for the whole country, but through this automatisisation the main meaning of the location of the coordinate was to be inside the unit.

The Address, Buildings and Dwellings Register is maintained by the municipalities. Within the application, there is a tool called GeoClient to set the coordinates for addresses and buildings. To make these coordinates meaningful for accessing and routing the rule was to set them close to the entrance within each building and address respectively. This worked well for the newly registered addresses and buildings, but the check of where the coordinates were put was not precise enough in the beginning and was not done for the initial data already in the system.

Over the years, these coordinates were used in a growing number of applications, both by the mapping agency and statistical office, but also by emergency services and for routing and planning purposes. It became obvious that the position of the coordinates was crucial, since it has an effect on the results. Routing applications often picked the wrong access road as e.g. the “calculate locations”-function snaps to the closest road section, even though this might not be the correct access road. So the location, precision and uniqueness of the coordinates became more and more important.

Routing problems before the change:

Navigation tools usually route to the closest point on the road network, even if the parcel cannot be accessed from this street.

Address coordinate of parcel number 126/1 at label point is closer to road on the left, but access is from the top road.	Coordinate is closer to point A (on wrong road) than B (on correct access road).	This is particularly problematic in allotments, where access to the whole set of houses often is only from one single entrance.
		

Solution

A milestone for the improvements of these coordinates came hand in hand with the launch of the Austrian Graph Integration Platform GIP; a joint and nationwide transport graph providing a multimodal digital map of Austria's transport network available to all authorities. This transport graph is more than just a street network and is a success story on its own.

It was decided to start a mass update of the address coordinates by “marrying” the Address register and the Graph Integration Platform GIP, meaning to make the address coordinates useful for navigation based on the GIP. Routing needs more than geocoded address points, it needs logical links to the street network.

In December 2016 the former address coordinates (then still often in the centre of an address) were moved automatically within the parcel but close (1m) to the GIP road section considered to be the correct access road and renamed to be the **address access points**. From the location of this address access point, the associated GIP coordinate was calculated automatically as the closest point directly on the associated road section.

Optionally the automatically calculated address access points can be improved manually by moving it to the driveway, which results in a recalculation of the associated GIP coordinate on the access road. The building coordinates were not affected by this process and stayed where they were, ideally near the entrance of the building.



Address coordinate of Bruckmühlgasse 2 (parcel number 126/1) at label point.

The parcel is on the corner of two streets, so it has two addresses, but Bruckmühlgasse 2 is the main address, since it is the street from which the parcel is accessed.



Bruckmühlgasse has the road ID 008146, which is also part of the official address, so it is included in both the address register as well as the GIP road section.

The address coordinate is moved perpendicular to the road section with the correct road ID and placed within 1m of the road, but inside the parcel. It becomes the new **address access point**.

The **GIP coordinate** is calculated as the closest point on this road section, so perpendicular and directly on the graph.



The building coordinate is not effected by this process and stays the same.

Remark: In this example it is placed in the centre of the building, so it does not correspond to the recommendation to place it near the entrance.



The address access point can be improved manually, e.g. moved to the actual driveway. The GIP coordinate is automatically updated.

Result

The new address access points are meaningful coordinates, useful for routing purposes as the last mile is routed correctly. Emergency services such as ambulance and fire brigade now reach the desired address faster.

The address access points replaced former address coordinates and are included in the address, building and dwelling register. To ensure the maintenance and sustainability of this project, the road graph GIP is also included in the new version of the GeoClient (the tool to set the coordinates within the address, buildings and dwellings register) as street layer.

As from the statistics point of view, the address access points are a great improvement and will be used in conjunction with the building coordinates for future routing applications. However, the access for Statistics Austria to the new GIP coordinates directly on the road section has not been clarified yet.

The address access points are available as zip-file from the mapping agency for free (two “as-of-dates” per year) subject to some licensing conditions. Download link:

http://www.bev.gv.at/portal/page?_pageid=713,2601271&_dad=portal&_schema=PORTAL

More information

- 1) Addressregister Guidelines (in German)
- 2) Information about the GIP (Graph Integration Platform): Presentation “GeoGIP - Adressen und GIP” © Dangl, Mandl-Mair, Rabl, Westhauser
- 3) Webpage about GIP, GIP.at and GIP.gv.at: <http://www.gip.gv.at/gipgvat-en.html>
- 4) Factsheet about GIP– The collaborative digital traffic network for all authorities:
http://www.gip.gv.at/downloads-219.html?file=tl_files/dynamic_dropdown/privateUploads/Downloads/Factsheets/Factsheet_Uebersicht_engl.pdf
- 5) GIP and GeoGIP – “marrying” the GIP with the Austrian address coordinates:
 - a) http://www.gip.gv.at/downloads-219.html?file=tl_files/dynamic_dropdown/privateUploads/Downloads/Praesentationen/AGIT17%20Mandl-Mair_Verein%20OeVDAT-die%20GIP%20im%20Vollbetrieb.pdf
 - b) http://www.gip.gv.at/downloads-219.html?file=tl_files/dynamic_dropdown/privateUploads/Downloads/Praesentationen/AGIT16_Unger%20Redl_Adressregister%20und%20GIP.pdf
 - c) http://www.gip.gv.at/downloads-220.html?file=tl_files/dynamic_dropdown/privateUploads/Downloads_engl/Presentations/AGIT14_GIPday_Redl_Zugangs-%20Geba%CC%88ude-%20und%20Grundstu%CC%88cksadresse_Umstzung%20in%20der%20VAO.pdf

Contact information

Ingrid Kaminger, Statistics Austria, Ingrid.kaminger@statistik.gv.at or geoinformation@statistik.gv.at

(C1.3) Building Points and the National Dwellings Register (Portugal)

Keywords: Principle 1, address register, building, national dwellings register, geospatial infrastructure, point-based foundation, surveys, statistics

This use case mainly refers to the recommendations provided within Principle 1, in particular the recommendations provided under the requirement 1.1, to use data from National Spatial Data Infrastructures and requirement 1.2, to use point-based location data for geocoding. It demonstrates how recommendation 1.1.2 is being implemented in Portugal.

Introduction

Until 2012, Statistics Portugal used a Master Sample (MS) to perform household surveys. After each Census and based on the information collected, a large sample of dwellings was designed, being maintained over the following decade through updates based on fieldwork. There were several reasons for Statistics Portugal to use such Master Sample, namely (i) legal issues (“Citizens shall not be given an all-purpose national identity number” – Article 35 of the Constitution); (ii) difficult to access administrative data sources; and, (iii) cost of data collection.

After 2012, aiming to improve the quality of official statistics and at the same time ensure the optimization, improvement, flexibility, modernization and efficiency of the statistical production process, Statistics Portugal started to use the National Dwellings Register (FNA) as the primary source of the Social Surveys.

The FNA contains information concerning identification (each unit as a unique identifier), type of dwelling (collective or conventional), occupancy status (usual residence, seasonal, vacant), tenure status, useful floor space and location, among others. The location is given by administrative division, NUTS, GRID, address and XY coordinates of the building (Census 2011).

Description of problem

In 2011 Statistics Portugal built a national geodatabase comprising all the georeferenced residential buildings from the 2011 Census (BGE). It is a point based nationwide coverage that is being continuously edited in an internal quality control process, and updated by the completed buildings and buildings permits (for new and/or demolished buildings), that include the XY location and addresses, that the municipalities provide to Statistics Portugal on a monthly basis.

The BGE is the spatial representation of the FNA buildings and, along with the GEOSTAT 1Km² GRID are the main geospatial data framework for the FNA, for the purpose of sampling and data collection.

The FNA consists of an exhaustive register of buildings and dwellings collected from the most recent Census (2011) which has been continuously updated from that moment onwards. Updating the FNA is a continuous process supported by information collected from fieldwork (dwellings selected for surveys) and the *Indicators System of Urban Operations*, (building permits), electronically filled by the municipalities, being it the main source of information on new buildings/dwellings and demolitions.

The FNA will be the main source for the Census 2021 operation, allowing to replace the traditional interviewer based distribution of questionnaires done in the field, by sending a letter through postal services containing the login details (access code and link) to complete the questionnaires online.

Challenge

In order to assure the maintenance and updating process of the FNA, Statistics Portugal had to

develop a geographic tool for the visualization by the field agents of the building and dwelling units from social surveys samples. The designed solutions was required to assure the navigation through the location of sampled buildings, provide management and control functionalities of the data collection process, and guarantee the possibility to correct the existing address data and geographical location when executing the fieldwork for the survey.

For this purpose, GeoINQ was developed as a Geographic Information Systems (GIS) WEB solution designed with the aim to integrate geospatial data into the production process of official statistics in an innovative way. It allows greater efficiency and rationalization of the resources especially in the household's surveys by supporting the data collection process.

GeoINQ is integrated, via webservices, with *Statistics Portugal Global Survey Management System* (SIGINQ-IE), and consumes a set of services and geographic datasets covered by the INSPIRE Directive.

SIGINQ-IE is Statistics Portugal surveys management system that integrates a set of subsystems (Meta-information system; Contacts Center - SICC; National Buildings and dwellings register file – FNA; Surveys by interview process management register system – GPIE-REG; Interviewers management system – ENTR; Sampling management system – SIGUA-UA)

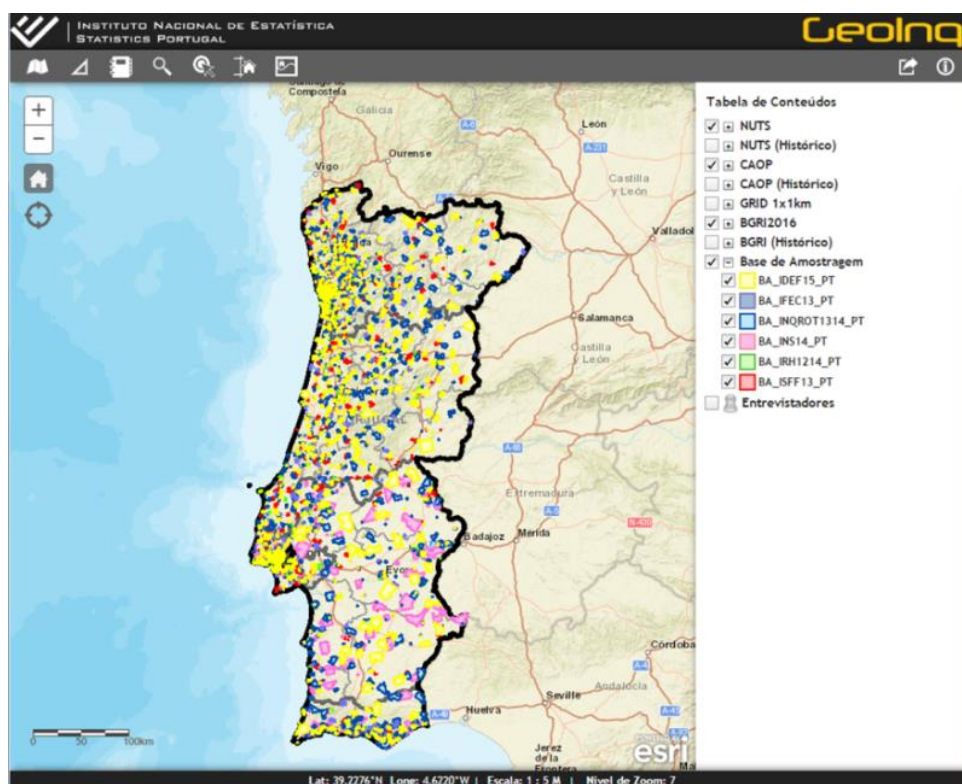


Figure 1: GeoINQ – Geospatial Data

GeoINQ contains a geographical framework with several geospatial feature layers like NUTS, Portugal Official Administrative Map of Portugal (CAOP), European 1km2 GRID, Census Statistical Units (BGRI), sampling frames (BA), Interviewers residence building and Base maps, including the Portuguese Mapping Agency (DGT) orthophotomaps.

The BGE contains the points (XY coordinates) of all the buildings georeferenced in the Census 2011 Operation updated through the above mentioned process.

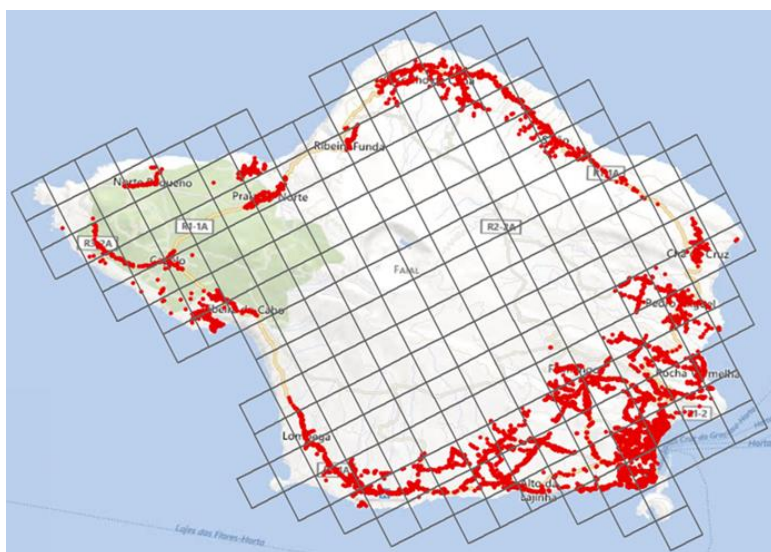


Figure 2: GeoINQ – Surveys samples

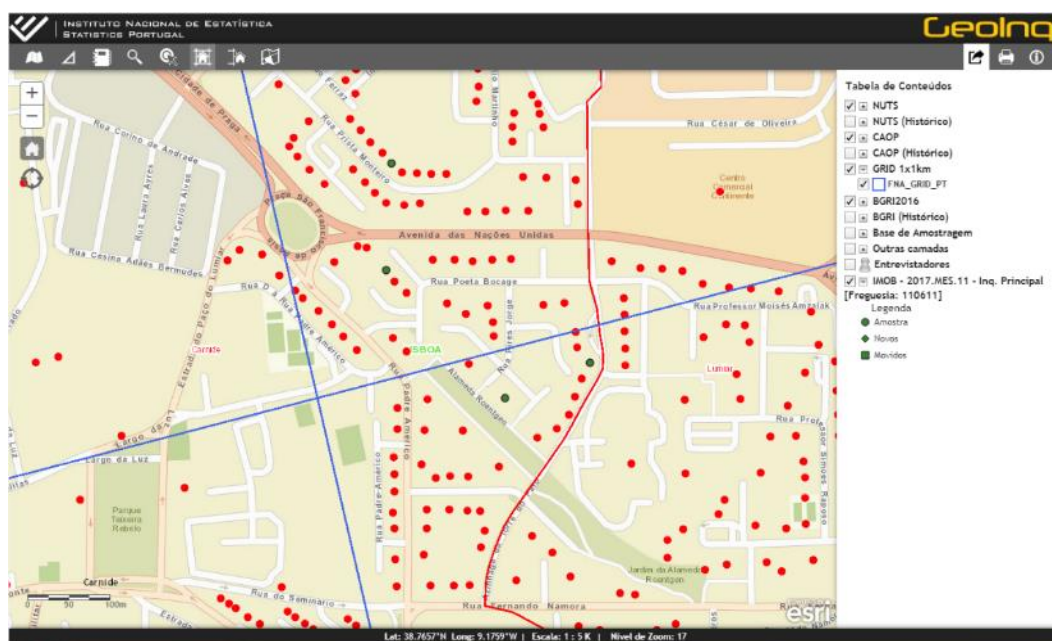


Figure 3: GeoINQ – Surveys samples

The main GeoINQ geospatial features is the ability to show the surveys samples assigned to interviewers - consisting of the FNA dwelling units materialized by BGE buildings points (points in green in the figure above).

GeoINQ – Functionalities

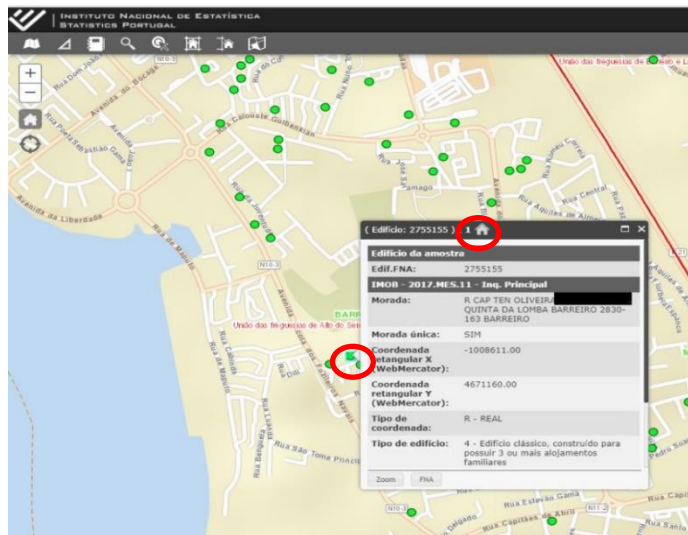
The GeoINQ app has a set of functionalities, supported in map services, which allow users to perform several geospatial operations – from the basic tools (zoom, search, query, etc.) to the more complex geoprocessing tools and creation of thematic maps.

- Search and visualize

GeoINQ users can apply spatial filters to search and visualize statistical units and samples from the different surveys. Users can only access the geographic and analysis information previously defined in their user profile. For example, only interviewers have permission to view the buildings of an entire sample from a specific survey that is assigned to them.

- Query

All users can query and obtain information about the attributes of BGE/FNA buildings that integrate social/household surveys samples and their respective FNA dwellings



- Edit

Interviewers that are in the field, collecting data in-person, have permission to edit BGE buildings geography and addresses directly in GeoINQ. With GeoINQ geoprocessing tools interviewers can move, add or delete BGE/FNA buildings.

GeoINQ is integrated with the FNA system and it also allows users to edit address attribute of FNA dwelling units.

Edits to the BGE buildings geography, done directly at GeoINQ, or to FNA buildings/dwellings address, done at FNA app via GeoINQ, proposed by interviewers are validated by the Geoinformation and FNA Management Units.

(Edifício: 1219229) | 0 (1 de 2)

Edifício da amostra

Edif.FNA: 1219229

Código censos 2011: 08110100102057

CENSOS - 2016 - Inq. Principal

Chave de Recolha (PSU): 08110100102057

Morada na amostra: TAPADA DA PENINA
PORTIMAO 8500-082 ALVOR

Edifício movido: Não

Tipo de coordenada: R- REAL

Estado de construção do edifício: 4- Edifício funcional

Tipo de edifício: 1- Edifício clássico, construído para possuir 1 ou 2 alojamentos familiares -

Zoom FNA FE Edifício **Mover**



Sim Não

FNA
Ficheiro Nacional de Alojamentos

V 1.4.0.0 de 20-10-2017

Proposta da morada de UA do FNA

Interviente: ext.margarida.sousa Perfil Interviente (cod): 0 Tipo de Utilizador: 0 Tipo de amostra (cod): 1 Operação Estatística (cod):

Tipo de Utilizador DMI: (S) Supervisor, (F) Supervisor Telefónico, (C) Codificador, (D) Codificador CATL, (R) Resp. Centro Rec./Reg. Aut. Presencial, (O) Resp. Centro Rec./Reg. Aut. Telefónico, (A) Resp. Centro Rec./Reg. Aut. CAVI, (I) Resp. Nacional CAVI, (N) Coord. Nacional Recolha e (P) Coord. Nacional Projecto

Inquérito: IMOB

Ocorrência: 2017.MES.11

Tipo de amostra: 1

Centro de recolha:

Perfil de interveniente:

PSU: 1504.B

Nº alojamento na PSU: F806

Código do Edifício: 2755155

Código do Alojamento: 025

Município: 1504 Barreiro

Número de entradas do edifício: 1

Resposta Urgente ☐

Incoerência entre localização do ponto no terreno e morada da UA ☐

Atualização toponímia da rua ☐

Tipo de via: 201103CENSO Rua

Designação de via: 201605II CAP TEN OLIVEIRA

Prefixo de edifício: 201103CENSO Torre

Designação de edifício: 201103CENSO 3

Número de porta:

Andar: 201103CENSO RC

Lado: 201103CENSO DTO

Sala:

Lugar: 201509II QUINTA DA LOMBA

Localidade: 201103CENSO BARREIRO

Código Postal: 201103CENSO 2830 163 BARREIRO

Observações:

Criar proposta

GeolNQ is a good example of geospatial data integration at several stages of the statistical production process, mainly in the data collection stage, contributing to a rationalisation of resources and a most efficient data collection process.

Contact information

Ana Santos, Statistics Portugal, ana.msantos@ine.pt or ine@ine.pt

(C1.4) A geocoded building and dwelling register as a base for the production of geospatial statistics (Switzerland)

Keywords: Principle 1, address register, building register, geospatial infrastructure, point-based foundation

This use case refers to most recommendations provided within Principle 1, to use fundamental geospatial infrastructure and geocoding of statistical information and Requirement 1.2, to use point-based location data for geocoding. It describes the flow of address data between agencies and the benefits of having one national, uniform and authoritative address register available for public institutions to include in their respective business processes. It also demonstrates the benefits of point-of-entry validation mechanisms and open address data.

Note: This case was not prepared for or by the GEOSTAT 3 project but is considered as good practice by the project group.

Introduction

The population census of 1990 was the first fully geocoded census in Switzerland. Since 2000, the geocoding process has been regularly improved and allows the geocoding of many statistics. During the last years, an ever closer cooperation was established between the Federal Statistical Office (FSO) and the Federal Office of Topography (Swisstopo), fostered by the federal law of geoinformation which provides optimal legal conditions. This cooperation is an important basis for a reliable and sustainable implementation of the national statistical geospatial framework.

The register of buildings and dwellings

The approach is mainly based on a standardized update of the register of buildings and dwellings (RBD) in a close collaboration with the land surveyors. The local construction authorities notify the RBD of all building projects, using a standardized data exchange procedure. New buildings are announced with basic location data, such as address, approximate coordinates and land plot number as well as descriptive data such as number of floors, type of heating, number of dwellings, rooms, etc.

During the registration process, the system provides a unique identification number to each building, each address and each dwelling. After that, the building information number is communicated to the local surveyors. After completion of the cadastral survey process, its results are communicated to Swisstopo, using a standardized interface. Finally, by accessing cadastral data the FSO can retrieve all necessary information for updating and validating the RBD data. It is then possible to publish an official list of addresses, including building and address identification number as a part of the national geodata infrastructure, on a free access basis. These identification numbers will be integrated in administrative and statistical datasets. Every dataset containing a building identification number can be geocoded easily during the statistic production process.

More information

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2017/mtg3/Paper_UNECE_v1_1.pdf

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2017/mtg3/S2_DOUARD_UNECE2017_CollaborativeApproach_v03.pdf

Contact information

Romain Douard, Federal Statistical Office of Switzerland, romain.douard@bfs.admin.ch or info@bfs.admin.ch

Rainer Humbel, Federal Statistical Office of Switzerland, Rainer.Humbel@bfs.admin.ch or info@bfs.admin.ch

(C1.5) The benefits of open address data (Denmark)

Keywords: Principle 1, address register, geospatial infrastructure, point-based foundation, open data

This use case mainly refers to the recommendations provided within Principle 1, that national geocoding services and address data should be open for other countries for cross-border geocoding purposes. It also illustrates how open address data can create strong incentives for the whole society to use and implement authoritative national data in their businesses as proposed within Principle 2, recommendation 2.5.4.

Note: This case was not prepared for or by the GEOSTAT 3 project but is considered as good practice by the project group.

Introduction

In 2002, the Danish government, having determined that “free and unrestricted access to addresses of high quality is beneficial to the public and forms the basis for reaping substantial benefits in public administration and in industry and commerce”, released its official Danish address database free of charge.

Conclusions

Eight years later, the government analysed the impact of opening up Danish address data and came to the following conclusion

- **Reuse:** In 2010, free-of-charge address data was delivered to total of 1,236 parties of which 70% were from private companies, 20% from the central government and 10% from municipalities.
- **Financial Benefit:** In 2009, an independent consulting firm determined that the direct benefit of free-of-charge address data from 2005-2009 was 62 Million Euros. This number was expected to rise in 2010 with the projected value of free-of-charge address data for 2010 projected to be 14 Million Euro. The total cost of the programme up until 2009 was around 2 Million Euros and was expected to be 0.2 Million euros in 2010.
- **Indirect or Derived Benefits:**
 - Elimination of Duplicate Collection: Releasing public address data has eliminated duplication of data collection.
 - Improved Public Service Coordination: Increased confidence that emergency services, ambulances, police and other emergency services all use the same reference data. An additional benefit is that, given that reporting errors or omission in the data has been simplified, consumers of the above-mentioned service can have increased confidence that the reference data is more accurate.
 - Higher Quality Data & Standardisation: Simplify the process by which errors and omissions are reported users can report by ensuring that errors only have to be corrected by one party, in one place. Furthermore, the release of free-of-charge address data has meant that there is now a standardised and known address data format for Denmark.

More information

<http://opendatahandbook.org/value-stories/en/danish-address-registry/>

<http://odimimpact.org/files/case-study-denmark.pdf>

(C1.6) The Real Property Register – authoritative location data for geocoding within the NSDI (Sweden)

Keywords: Principle 1, location data, addresses, buildings, cadastral parcels, geospatial infrastructure, point-based foundation

This use case mainly refers to the requirements provided within Principle 1, to use authoritative data from NSDIs and to use point-based location data for geocoding. In particular, it demonstrates how different geospatial objects (cadastral parcels, buildings and addresses) can be hierarchically and consistently linked to each other, to form one national and uniform geocoding infrastructure.

Introduction

For decades, the Swedish National Real Property Register has been used by Statistics Sweden and other public institutions to geocode statistical and administrative information. The register has deep historical roots and has evolved gradually over the years. As of today, it comprises a comprehensive repository of authoritative location data available within the NSDI, enabling flexible geocoding of information to the level of cadastral parcels, buildings and addresses. The information in the Real Property Register is national location masterdata and feeds into a number of other public institutions such as Statistics Sweden, Tax Administration etc. Figure 1 below illustrates the flow of address information (which is part of the Real Property Register).

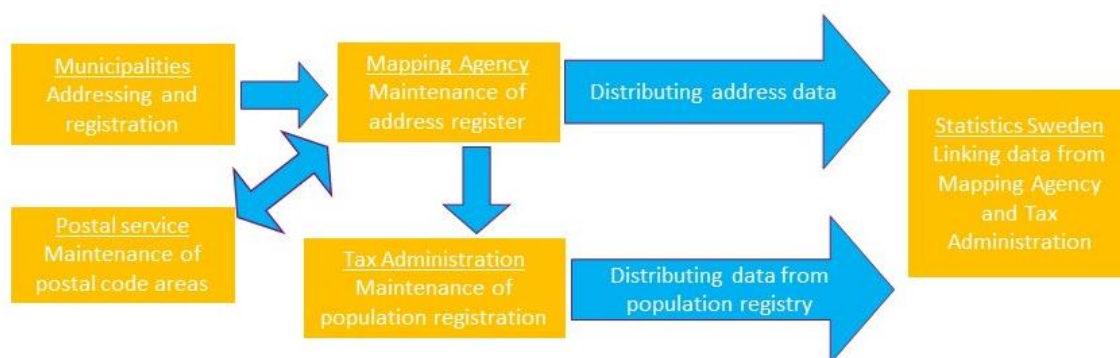


Figure 1: Illustration of data flow for address information

A compound data repository

To most users, the Real Property Register appears as one single collection of data concerning all properties. Despite a homogenous and concise presentation of the textual information, the register is a compound of five parts maintained by different authorities:

- (1) The general part (cadastre) including the cadastral index map

In Sweden, the word 'cadastre' is seldom used. However, the general part of the Real Property Register corresponds well to the international meaning of that term, as it is the official register of the country's division into property units or cadastral parcels. For all such units, it contains information about property designation, estimated acreage and location (by a pair of coordinates representing the centre of the parcel).

The cadastral index map is a geographical representation of all current properties, showing their unique designations and approximate boundaries at an original scale of 1:10,000 or 1:2,000 (rural

and urban areas respectively). Besides the basic “cadastral parcel layer”, other topographic details, buildings and land use features are also included in the map.

The general part of the Real Property Register, as well as the cadastral index map, is maintained by the National Mapping Agency (Lantmäteriet).

(2) The land register part

Land registration is mainly focusing on real property ownership and other kinds of rights created outside cadastral matters, e.g. leases and mortgages. After property conveyances or other transactions, the titles are registered and thereby secured against the third party.

(3) The address part

Properties are recorded with physical addresses in the Real Property Register. Addressing is under the responsibility of the municipalities. However, address data are registered by the municipalities directly into the central repository. The address part of the Real Property Register is the one and only authoritative address register in Sweden.

(4) The building part

The building part contains basic data concerning buildings on each property, e.g. what type of buildings they are (dwellings, non-residential premises, industries etc.). Every building has a unique identity and can be located geographically through its centre coordinates. Each building is also represented as a polygon feature in the cadastral index map with the same unique identity found in the register. These data are maintained by the municipalities and registered directly into the central repository.

(5) The property tax assessment part

The National Tax Authority is responsible for estimating a general tax assessment value for most properties. In order to do so, they use assessment models for mass valuation, including land value maps, provided by Lantmäteriet. Statistics of property sales, registered in the land registry part of the Real Property Register, form the basis for these models.

(6) Dwellings

After a legislation in 2006 on establishment of a dwelling register, it was incorporated as a part of the Real Property Register and implemented in 2010. The dwelling register is a correspondence table linking addresses and buildings together with additional information about individual dwellings (dwelling ID). The dwelling register is used by the Tax Administration for population registration. The dwelling register does not have a spatial representation of its own, but can be linked to buildings and/or addresses.

Integrated location data objects

The geospatial objects of the Real Property Register (cadastral parcels, buildings and addresses) are consistent and hierarchically linked to each other, both conceptually and topologically, which enables inclusion of all three object-types in the geocoding infrastructure.

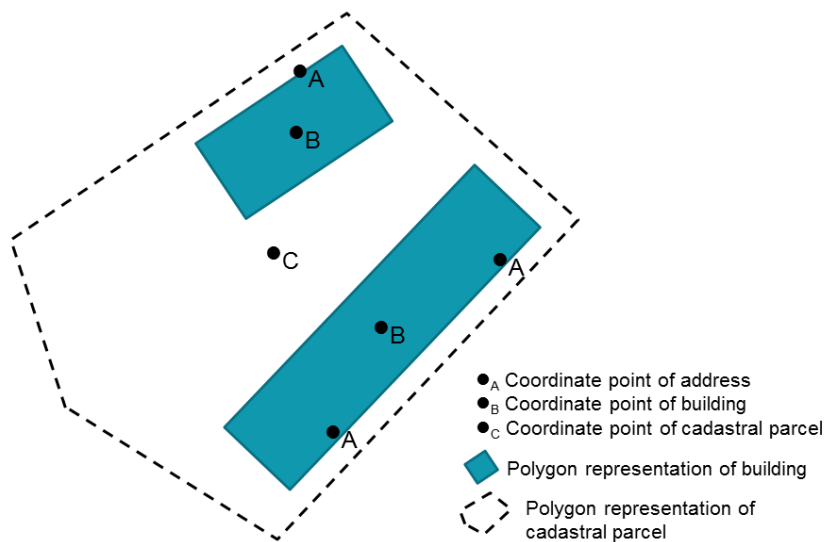


Figure 2: Conceptual illustration of the integrated and hierarchical location data framework of the Real Property Register.

As shown in the figure above, coordinates of buildings (B) and addresses (A) are linked to the cadastral parcel in which they are located. The cadastral parcel can be spatially represented by its centroid coordinate (C) or by a polygon feature from the Cadastral index map representing its extent. The coordinate of an address (A) is linked to the building (B) to which it belongs (typically entrance point). A dwelling does not have a spatial representation of its own, but can be linked to building and/or address location.

More information

<https://www.lantmateriet.se/en/real-property/Fastighetsinformation/Fastighetsregistret/>

http://www.theboundary.no/ep_tmp/files/204504289849f0e16a0ccaf.pdf

Contact information

Jerker Moström, Statistics Sweden, jerker.mostrom@scb.se or scb@scb.se

(C1.7) Building and Address data – validation process in cooperation with the Municipalities (Portugal)

Keywords: Principle 1, address register, buildings, national dwellings register, open data, point-based foundation, municipalities

This use case mainly refers to the recommendations provided within Principle 1, particularly the recommendations provided under the requirement 1.1 to use data from National Spatial Data Infrastructures and requirement 1.2 to use point-based location data for geocoding and more specifically recommendation 1.1.2 and recommendation 1.1.3, on systematic management of unique identifiers.

Introduction

The use of cartography supports data collection at Statistics Portugal since 1981. In 1995, Statistics Portugal started the preparation of the cartographical infrastructure to support the 2001 Census, named as “Geographic Information Referencing Base” (BGRI 2001) corresponding to the smallest census areas - statistical units, which was based on a Geographic Information System (GIS).

With the thematic Spatial Data Infrastructure (SDI) development and the creation after the 2011 census round of a point based foundation spatial framework - supported by the Geographical Building Base (BGE), Statistics Portugal has been increasing the integration of the spatial component at several stages of the statistical production process and managing the INSPIRE national thematic datasets under its responsibility.

At the national level, Portugal has structural failures due to the absence of a georeferenced land property registry and a system for its management, as well as the inexistence of a national wide official coverage for addresses and buildings registers.

At the local level, the Portuguese municipalities are one of the main producers of georeferenced detailed data and one of the entities that are involved in the implementation of INSPIRE within the National Mapping working groups. Within the National Statistical System municipalities provide to Statistics Portugal, on a mandatory monthly basis, data regarding completed buildings and building permits (for new and/or demolished buildings), that include the XY location and addresses. They are also one of the major contributors of open geospatial data for the Opens Street Map (OSM) platform.

Description of problem

Since the 1991 and in every ten years, Statistics Portugal implements a specific program to update the statistical units’ boundaries, in connection with the pre-enumeration censuses phase.

Traditionally the strategy of Statistics Portugal was based on performing intensive internal editing work and also on strong field collaboration with the municipalities.

The continuous editing work performed since 2011 by Statistics Portugal over the buildings dataset along with the great achievements obtained by the geospatial community, in particularly at local level, and the increasingly available open datasets, showed the need to define a new update geospatial data strategy.

Challenge

For the 2021 census round, Statistics Portugal needs to draw up a plan to update the INSPIRE compliant Statistical Units, Named Places and Buildings datasets.

This strategy must promote the use of building and address data available in the municipalities that should be aligned with the work to be performed by them regarding the scenario of INSPIRE implementation in a process conducted by the Portuguese NMCA – Directorate General of the Territory.

Statistics Portugal will begin the development of a National Data Infrastructure (IND) in 2019, which will make available a set of information and resources from a single point of entry, based on better statistical information, greater analytical capacity and flexibility of adequacy of information to the needs of decision making. The above mentioned strategy should ensure the implementation of the space component of the IND, in order to guarantee the geointegration of the administrative databases and intensify the processes of administrative data appropriation for statistical purposes.

To apply this strategy a work program of cooperation between Statistics Portugal, the municipalities and the commonwealth (CIM) as a voluntary associations of Portuguese municipalities, will be launched in the first quarter of 2019.

A technological solution and data editing methodology were specified to guarantee:

- the creation of common unique identifiers and the interoperability between the Statistics Portugal geospatial databases and the municipal geographical bases
- the use of structured procedures, harmonized and documented in a clear example of good practices

Statistics Portugal expect that this cooperation program will contribute for the development of national building and address datasets, able to integrate the Portuguese National Spatial Data Infrastructure (NSDI) and relevant to the geocoding of statistical information in the way that is required by Principle P1 of GSGF.

Contact information

Ana Santos, Statistics Portugal, ana.msantos@ine.pt or ine@ine.pt

2 Principle 2: Geocoded unit record data in a data management environment



(C2.1) Geocoding guidance prepared by ABS (Australia)

Keywords: Principle 2, geocoding guidelines, national geocoding standard

This use case particularly refers to recommendation 2.3.1, provided within Principle 2, stating that Member States should develop and apply national guidelines for geocoding workflows in order to ensure a consistent and conform result within and between institutions. The case gives a concrete example on what such guidelines may comprise.

Note: This case was not prepared for or by the GEOSTAT 3 project but is considered as good practice by the project group.

Introduction

The Australian Bureau of Statistics (ABS) has prepared guidance material to provide national guidelines for geocoding of statistical information. The guidelines cover three of the main options for geocoding of unit record data in socio-economic datasets. The guidelines describe the basic elements and processes applied when implementing these three options, and provides references to resources associated with them. The geocoding standards and guidelines are part of ABS' Statistical Geospatial Framework guidance material.

More information

[http://www.abs.gov.au/websitedbs/d3310114.nsf/home/Statistical+Spatial+Framework+Guidance+Material/\\$File/Geocoding+Unit+Record+Data.pdf](http://www.abs.gov.au/websitedbs/d3310114.nsf/home/Statistical+Spatial+Framework+Guidance+Material/$File/Geocoding+Unit+Record+Data.pdf)

<http://www.abs.gov.au/websitedbs/d3310114.nsf/home/Statistical+Spatial+Framework+Guidance+Material>

Contact information

Australian Bureau of Statistics: <http://www.abs.gov.au/>

(C2.2) Dutch Geocoding service API (Netherlands)

Keywords: Principle 2, Principle 5, geocoding service, point-based geocoding, machine-readable data, API

This use case mainly refers to the requirement provided within Principle 2 to store location only once, in this case by using a geocoding service API (recommendation 2.2.3) set up by the Dutch Cadaster. In addition, it demonstrates a practical implementation of the recommendation within Principle 4, to set up geospatial services in a service-oriented architecture to standardise geospatial production components (4.1.6).

Introduction

The Dutch “Locatieserver” is a good practice example of a geocoding service API requested by the GSGF principle 2. It is hosted by the Dutch hosting organization PDOK organized by the Dutch Cadaster. The API will return different type of objects based on search strings. It can return an object-id from the Dutch Base registry for Addresses and Buildings (BAG) and in the future, it will return also other feature types like Statistical Units (neighborhoods and districts)

It works for one search at a time, but because it is developed as an API, it can be used in programs that can geocode a whole list of addresses.

A Dutch description can be found at:

<https://github.com/PDOK/locatieserver/wiki/API-Locatieserver>

The basic URL for using the API is:

<https://geodata.nationaalgeoregister.nl/locatieserver/v3/suggest?q=<search string>>

It distinguishes several parameters like explained below:

q=<search string>

Example q="Gouda"

Optional: fl=<filed name>

Example fl=centroide_ll (will only give the Lat Long coordinates)

Optional: sort=<sort method>

"score desc", "sortering asc", "weergavenaam asc"

Optional: rows=<number of results>

default = 10

Optional: start=<index>

Useful when you want to slit up search results. Default-value is "0".

Optional: wt=<format>

"json" or "xml". Default-value is "json".

Optional: indent=<value> "true" or "false".

To define the indent in json

Optional: lat=<latitude>&lon=<longitude>

Example: lat=52.09&lon=5.12

Results will be sorted in distance to this point

Optional: fq=<filter query>

Example fq=bron:BAG or fq=type:adres

An Example of a queries can be:

https://geodata.nationaalgeoregister.nl/locatieserver/v3/suggest?q=gouda&fq=type:gemeente&wt=xml&fl=centroide*

This will return:

```
<?xml version="1.0" encoding="UTF-8"?>
- <response>
  - <result maxScore="23.71952" start="0" numFound="1" name="response">
    - <doc>
      <str name="centroide_ll">POINT(4.70620369 52.0153015)</str>
      <str name="centroide_rd">POINT(108250.851 447657.184)</str>
    </doc>
  </result>
  - <lst name="highlighting">
    - <lst name="gem-c9c42cc32725bb507a2931c210877dcb">
      - <arr name="suggest">
        <str>Gemeente <b>Gouda</b></str>
      </arr>
    </lst>
  </lst>
  - <lst name="spellcheck">
    - <lst name="suggestions">
      - <lst name="gouda">
        <int name="numFound">4</int>
        <int name="startOffset">0</int>
        <int name="endOffset">5</int>
        - <arr name="suggestion">
          <str>golda</str>
          <str>goudb</str>
          <str>goude</str>
          <str>gouds</str>
        </arr>
      </lst>
    </lst>
  </lst>
  - <lst name="collations"/>
</lst>
</response>
```

Contact information

Ir. Pieter Bresters, Statistics Netherlands, p.bresters@cbs.nl or info@cbs.nl

Drs. Niek van Leeuwen, Statistics Netherlands, n.vanleeuwen@cbs.nl or info@cbs.nl

(C2.3) National Geocoding and Routing Services (Finland)

Keywords: Principle 2, geocoding service, point-based geocoding, API, machine-readable data, routing

This use case mainly refers to Principle 2 covering, in wide sense, all its requirements. In order to be realised in practice it also refers to principle 1 (especially to requirement 1.1). In addition, it demonstrates a practical implementation of the recommendation within Principle 4, to set up geospatial services in a service-oriented architecture to standardise geospatial production components (4.1.6).

Introduction

Statistics Finland has several needs related to addresses: converting them to coordinates (geocoding), converting coordinates to addresses (reverse geocoding) and calculating distances between two addresses. Over time, several ad-hoc solutions have been developed for geocoding and there is a need to replace them with one single service.

Distances between two addresses are largely calculated as the crow flies even if it was about people's ways to work. Calculations made along transport networks and including travel time are desired for better precision. Other authorities in Finland have developed public services for these purposes. Statistics Finland aims to utilise them and also to contribute to their development.

Description of the problem

Over time, the various needs for geocoding have resulted in several different custom applications within Statistics Finland. Besides that, desktop GIS applications and even Google are used. Results from different applications are not uniform or reproducible by others and maintenance causes unnecessary work.

Solution

The National Land Survey of Finland is developing a national service for geocoding and reverse geocoding as a part of their Geospatial Platform Project. It also aims to create a National Address System which these services will be using as a data source. Statistics Finland will participate in the project to ensure that it also satisfies the needs of statistical use such as returning the permanent id of buildings along with addresses or coordinates. This is done by either contributing to the core service or by making a custom service of our own on top of it. Routing will be based on the routing API Digitransit.fi provided by the Finnish Transport Agency and the Helsinki Region Transport Agency. It allows querying routes between street addresses at a given time using selected means of transportation from walking or cycling to taking an airplane. A service customised for the needs of statistical use will be made on top of this API by Statistics Finland.

The good practice here is not something that can be implemented by a single agency itself. It calls for an open and cooperative national spatial data infrastructure, which offers the basic building blocks for developing custom solutions.

Result

The solution is at the planning stage. The Digitransit.fi API for routing is already in public use and the National Land Survey of Finland's API for geocoding is being tested and scheduled for release in February 2019.

More information

<http://www.paikkatietoalusta.fi/en>

<https://digitransit.fi/en/developers/>

Contact information

Tapio Kytö, Statistics Finland, tapio.kyto@stat.fi or info@tilastokeskus.fi

(C2.4) Geocoding Quality Declaration at Object Level (Norway)

Keywords: Principle 2, point-based geocoding, geocoding metadata, quality declaration

This use case refers to the requirements provided within principle 2, to ensure consistency and quality of geocoding results. In particular, it demonstrates a practical implementation of recommendation 2.3.3, that geocoding metadata should be provided at object level so that the accuracy of the assigned location can be assessed for each observation. The case also illustrates some aspect of requirement 2.1 to build an effective and secure data management environment.

Introduction

Statistics Norway (SN) receives a copy of the Official Business Register every night. A challenge to use this in GIS analyses is the lack of X- and Y-coordinates. As the conceptual model shows, there is an Address Control B for the Business register against the official Cadastre, which holds the official addresses, buildings and properties in Norway. Unfortunately, this control is only a check of right spelling of street names and does not return a unique numeric address code for the exact address. Due to this fact, SN must do an in-house geocoding process to obtain coordinates for each establishment.

The Address Control A is returning the unique numeric address – see Address terms below. The population of establishments geocoded must have location in Norway.

The registers used in the geocoding processes

The Business Register (BR) in SN is a copy of the Official Business Register mentioned above as Enterprises and Establishments. The aim of the geocoding processes is to attach X- and Y-coordinates to each Establishment (Local Kind of Activity Unit – LKAU) in the register.

The registers used for finding coordinates are all copies of Norwegian official administrative registers.

The Address Register

The copy of the official Cadastre holds both addresses, buildings and properties. For short, this copy is referred to as the Address register in this document. The basic statistical unit is also a part of this register. This register is used in the joining processes **A1, A2, A3, A4, A5, L1_b, M1, N1, V1, W1, W2, W3, W4** and **W5**.

The Population Register

This is a register of all citizens in Norway. This register is used in the joining processes **R1 and R2**.

The Norwegian Farm Register

The register includes all land and forest properties in Norway. Statistics Norway get a copy of this register twice a year. This register is used in the joining process **L1_a**.

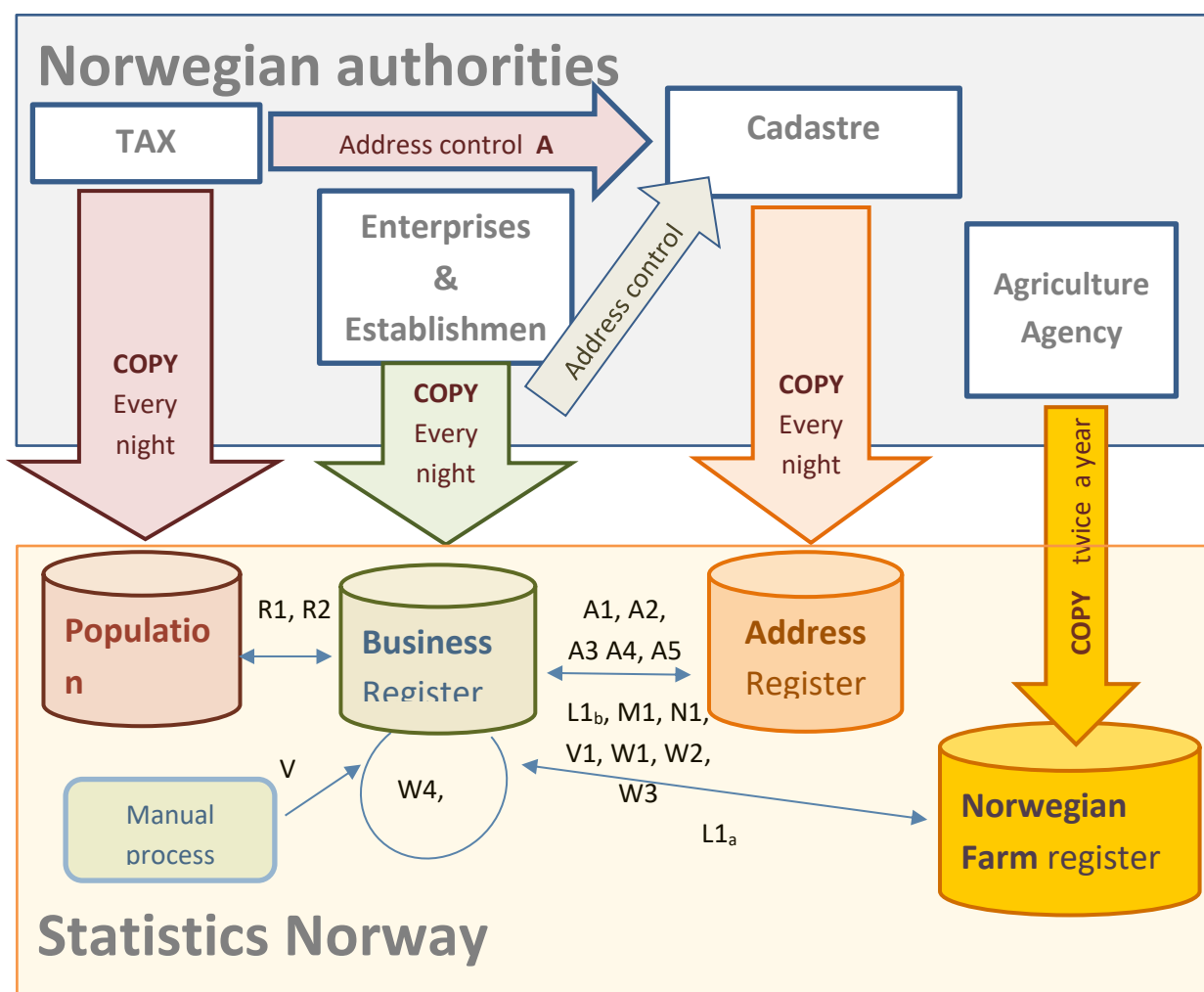


Figure 1: Conceptual model for geocoding the Business Register in Statistics Norway

Address terms

Norway has a coherent and centrally maintained address system that consists of two different address objects; street address and cadastral address. They correspond to unique numeric addresses as shown in the table below.

Part of the street addresses	Characters in the unique numeric address	Part of the cadastral addresses
Municipality number	- 4	Municipality number
Road number	5 - 9	Cadastral unit number
House number	10 - 13	Property unit number
Letter	14 - 17	Leasehold number
Condominium number (0000 for street addresses)	18 - 21	Condominium number
Apartment number	22 - 26	Apartment number

Depending of the need for accuracy, the number of characters used when joining registers may differ. When the purpose to geocode exact X- and Y coordinates, the first 17 characters in the unique address are used. Basic Statistical units are not part of the unique address.

Modified address is a term used in this document. This is street names spelled in a uniform way based on the original addresses attempting to adapt to street names in the Address register. Example is "Skolevegen" and "Skoleveien" (School street) where "vegen" and "veien" both are

correct ways of spelling the word for *street* in Norwegian. Other ways of modifying street names are reduction of blank characters, hyphens and periods. Expansion of abbreviations to the relevant word is also a way of modifying street names. Example of the latter could be “gt” expanded to “gate” which is the Norwegian word for *street* or *road*.

The joining processes

The joining processes runs every month, and the processes that provide coordinates are A1, A2, A3, A4, A5, L1, M1, N1, R1, R2, V1 and V2. These processes return also *basic statistical units* in most cases. The rest of the joining processes provide basic statistical units. The joining codes are given in a combined alphabetical and numerical order. I.e. the joining code A1 is likely the most correct address and the code W5 is the most uncertain address. Joining code XX means the establishment did not match at all.

Results from the joining in October 2017:

Joining process	Number of joins	Per cent
A1	491978	80.97
A2	787	0.13
A3	1508	0.25
A4	9465	1.56
A5	424	0.07
L1	19440	3.20
M1	6	0.00
N1	1186	0.20
R1	22876	3.76
R2	2572	0.42
V1	2049	0.34
V2	4	0.00
W1	132	0.02
W2	7165	1.18
W3	910	0.15
W4	3271	0.54
W5	43079	7.09
XX	772	0.13
Total	607624	100

Results from the joining in June 2018:

Joining process	Number of joins	Per cent
A1	499581	82.51
A2	578	0.10
A3	821	0.14
A4	9260	1.53
A5	272	0.04
L1	17937	2.96
M1	6	0.00
N1	1115	0.18
R1	20467	3.38
R2	2325	0.38
V1	1595	0.26
V2	4	0.00
W1	114	0.02
W2	6707	1.11
W3	974	0.16
W4	2947	0.49
W5	39948	6.60
XX	844	0.14
Total	605495	100

These results show slightly increase for the best joining criteria – A1. For the rest the results are on the same level.

Address variables used in the different joining processes and what is returned**A1**

Address terms in the join	Returns
Municipality number Street name House or property unit number Letter or leasehold number Post code	Coordinates and basic statistical unit

A2

Address terms in the join	Returns
Municipality number Modified street name House or property unit number Letter or leasehold number Post code	Coordinates and basic statistical unit

A3

Address terms in the join	Returns
Municipality number Modified street name House or property unit number Letter or leasehold number	Coordinates and basic statistical unit

A4

Address terms in the join	Returns
Municipality number Modified street name House or property unit number Post code	Coordinates and basic statistical unit

A5

Address terms in the join	Returns
Municipality number Modified street name House or property unit number Letter or leasehold number Post code	Coordinates and basic statistical unit

L1

This is a 2-step joining – called L1_a and L1_b in the conceptual model above

L1 _a - Variable used in the join	Returns
Organisation number	Municipality number Cadastral number Property unit number Leasehold number Post code
L1 _b - Address terms in the join	Returns
Municipality number Cadastral number Property unit number Leasehold number Post code	Coordinates and basic statistical unit

M1

Address terms in the join - (manually coded in BR)	Returns
Municipality number Cadastral number Property unit number Leasehold number	Coordinates and basic statistical unit

N1

Joining description	Returns
Some industry codes in the Business Register are matched to building type in the address register and when it is only one match within one post code the coordinates and basic statistical unit are returned. Example of building types are schools, nursing homes and power stations.	Coordinates and basic statistical unit

R1

Joining description	Returns
For one-man companies in certain industries having the same municipality and post code as in the owners private address the coordinates and basic statistical unit are returned.	Coordinates and basic statistical unit

R2

Joining description	Returns
One-man companies in certain industries gets the owners private address from the Population register.	Coordinates and basic statistical unit

V1

Joining description	Returns
Phonetic geocoding of modified addresses in BR against Address register	Coordinates and basic statistical unit

V2

Joining description	Returns
Manual coding of coordinates for Establishments around Oslo Airport while awaiting addresses for this area in the official Cadastre.	Coordinates

W1

Joining description	Returns
Addresses in BR joined against outdated addresses without coordinates in the Address register	Basic statistical unit

W2

Joining description	Returns
Joining by municipality number and modified street name. Requires streets to be within one and only one basic statistical unit.	Basic statistical unit

W3

Joining description	Returns
Joining by municipality number, alternative address (name of place) and post code.	Basic statistical unit

W4

Joining description within BR	Returns
If the Establishment has the same municipality and post code as one or more of the Establishments joined in L1 , the basic statistical unit as the majorities of the Establishments is returned.	Basic statistical unit

W5

Joining description within BR	Returns
The basic statistical unit returned is the one that have most businesses within the same post code.	Estimated basic statistical unit

XX

Businesses not having any kind of geocoding get XX in joining code.

Contact information

Marianne Dysterud, Statistics Norway, marianne.dysterud@ssb.no or ssb@ssb.no

(C2.5) Point-of-entry validation in data collection (Estonia)

Keywords: Principle 2, point-based geocoding, point-of-entry validation, population registry, address register

This use case mainly refers to the recommendations provided within Principle 2, Geocoded unit record data in a data management environment, particularly requirement 2.5 to use point-of entry validation in collection of administrative or statistical data. The use case shows how using the gazetteer improves the address data quality in Administrative Population Registry. Using of the gazetteer becomes available after implementation of address standard as part of the fundamental geospatial infrastructure (Principle 1).

Introduction

In order to improve the quality of Estonian address data used in different registers, an Address Data System (ADS) was implemented. With this system, Estonian addresses are standardised, which improves their quality in geocoding and the linking to different registers. Changes in the registers are made using a gazetteer, which makes the new address compliant with the standard.

Description of problem

In the preparation of the 2010 census, it was noticed that the Estonian address system needed quality improvements. During data entry in the registers, there were problems with typing, as one name could have different spellings; and problems with territorial hierarchy, as a settlement could be marked under a neighbouring municipality or county. In several regions, the share of automatically geocoded persons was less than 80 percent.

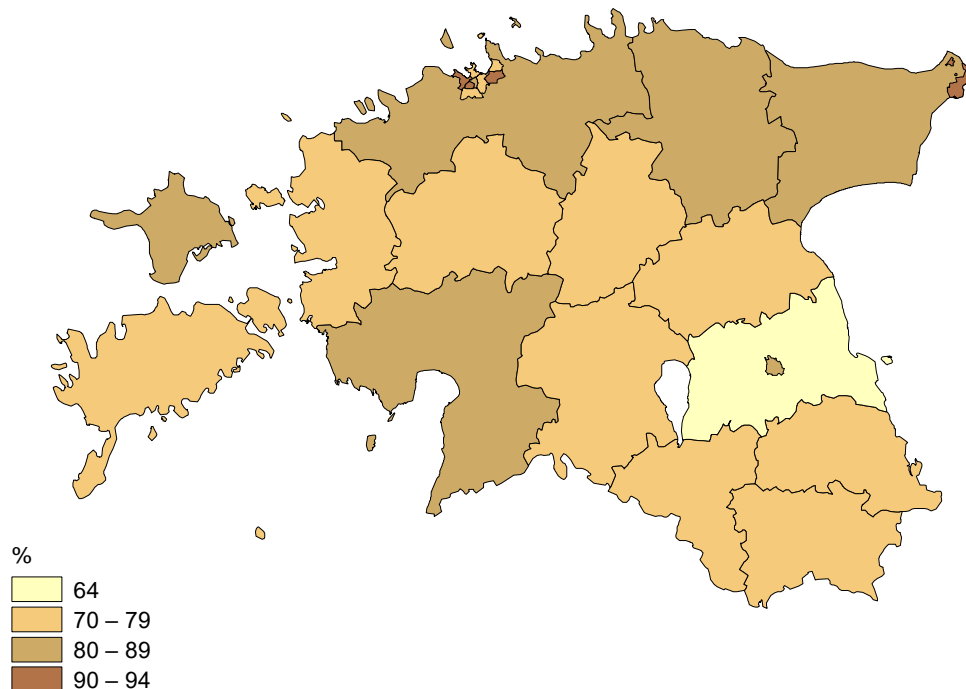


Figure 1: Share of automatically geocoded persons, before ADS implementation, 2008

Solution

ADS

The Address Data System (ADS) was adopted in 2008. The system has a register status, and its main goals are the creation of a central database of addresses and the implementation of a standard for geographical addresses.

The ADS coordinates co-operation between the administrative bodies involved and ensures the maintenance of databases containing address data. According to the ADS regulation, all databases belonging to the State Information System should use the ADS Management System for data processing.

In the Estonian Address Data System, an address consists of 8 levels/components (Figure 2). Geocoding means that the address is linked to an object in the nature. It has been assigned an address object identifier and is, therefore, also coordinated.

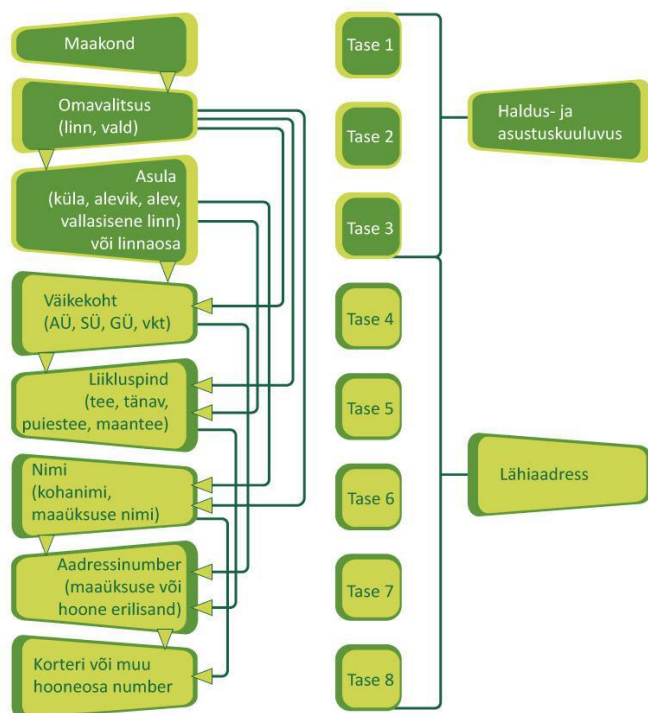


Figure 2: ADS structural elements or components

Official address data will reach the ADS mainly through the address data processing application and the State Register of Construction Works. Data are submitted by local governments. Data of traffic areas and small places are submitted through the Place Names Register. Data for administrative and settlement units are submitted through the Land Cadastre.

As the chief processor of ADS data, the Estonian Land Board has developed many ways for the normalisation and geocoding of addresses:

- 1) X-Road service,
- 2) mass geocoding from the table in the online application,
- 3) geocoding of a single item in the public online application.

All these possibilities currently only function with valid ADS addresses. The historic state is not taken into account, but the Estonian Land Board has noted it as a development need.

Linking statistical registers to ADS

Administrative registers have linked to the ADS using X-Road service. The administrative Population Register has fully implemented the ADS standards. Most of the addresses have been supplied with an ADS address identifier and an address object identifier. The Business Register only recently linked their information system to the ADS.

Changing addresses using gazetteer

After changing their place of residence, people can report the new address to the Administrative Population Register using a web application, which allows selecting the address components from a gazetteer list (Figure 3). If the address components are not found, it allows writing the address or a part of it as text. For example, if the street name is not found in the list, the following levels cannot be chosen and have to be written as text values.

UUE ELUKOHA ADDRESSI ANDMED:

Uue elukoha asukoht: Eesti

UUE ELUKOHA ADDRESS EESTIS: (kui aadress puudub nimekirjas, siis märkige puudulev komponent lahtrisse „Tekstiline aadress“)

Riik	Eesti
Maakond	Harju maakond
Vald/linn	Tallinn
Asula/linnaosa	Haabersti linnaosa
Väikekoht	Tühi
Tänav	Vali...
Aadressobjekti nimi/talu	Vali...
Maja	Vali...
Korter	
Tekstiline aadress	
Postiindeks	

Uue sideaadressi andmed: (sideaadressina esitatakse põhilukohast ERINEV aadress, juhul kui ajutiselt elatakse pikemat aega mõnes muus kohas)

Figure 3: Web Application for reporting address changes to Administrative Population Register

The desktop of local authority officials, who can also report address changes to the Administrative Population Register, uses a similar service, but there is no possibility to insert address as text. When the required address components are not available in the ADS, a local authority representative will

inform the Land Board and arrange an address. Then the Administrative Population Register will receive the correct address during the ADS–Administrative Population Register data exchange.

Result

After implementation of the Estonian Address Data System, the address data entry in the registers will be performed by using gazetteers. Now the share of automatically geocoded persons is over 97 percent in most regions (Figure 4).

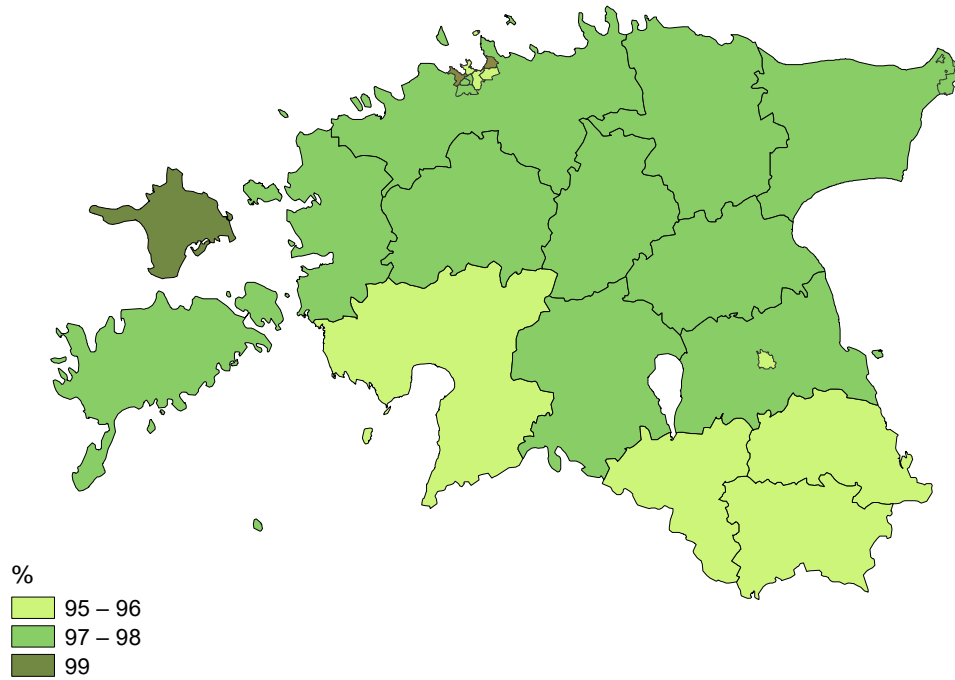


Figure 4: Share of automatically geocoded persons after ADS implementation, 2017

More information

Public service of the Address Data System: <http://xgis.maaamet.ee/adsavalik/ads>

Contact information

Ülle Valgma, Statistics Estonia, ulle.valgma@stat.ee or stat@stat.ee

3 Principle 3: Common geographies for production and dissemination of statistics



(C3.1) REGINA – Management of coding system for administrative geographies (Sweden)

Keywords: Principle 3, management of statistical and administrative geographies

This case mainly refers to Principle 3 and requirement 3.1, to set up and maintain a consistent framework of national statistical and administrative geographies. It demonstrates a solution for maintaining coding systems for administrative geographies and tracking changes over time. It also exemplifies how this information can be made available for users.

Introduction

Statistics Sweden is responsible for maintaining a number of administrative and statistical geographies, including maintenance of numeric codes that identifies each region, in addition to the names. Statistics Sweden is not responsible for assigning codes or names. Several authorities and organisations are responsible for specific types of divisions. However, when decisions are taken it is under the responsibility of Statistics Sweden to coordinate the result. Statistics Sweden must also disseminate the result, by informing the public about the changes. Both current and historical divisions must be available. The coding system is fully harmonised with the Cadaster maintained by the Swedish Mapping Agency. Any decision on territorial changes occurring during a year enter into force as of first of January the following year.

In order to make the system more accessible, Statistics Sweden has created a web service, to enable search for a division's code, name or geography, displayed as current state or as changes through the years.

Description of problem

Statistics Sweden's database on administrative geographies contains information on numeric codes and names for each unit. Any changes to administrative geographies, such as a merging or splitting of municipalities or transfer of land areas between municipalities, are recorded in the database along with information about the dates when the changes have occurred. The database holds information on any changes all the way back to 1952, when the modern administrative system based on municipalities were established. The data can be combined with other registers in order to create statistics on regional or smaller areas.

As the access to the database used to be restricted, Statistics Sweden got many inquiries from users, about codes, names or geographic division of a certain area. Several of the questions required a tailor made extract from the database. There were about 80-110 requests per year that involved extraction and delivery of tailor made data.

The information was also published at Statistics Sweden's website in MS Excel and PDF format with long lists of current counties, municipalities and parishes. There were also lists of the latest changes in codes and names.

The spectrum of users comprises other public bodies, researchers, historians, students, genealogists and others who need information about administrative divisions and their changes over time. Statistics Sweden itself is also a major user of this information.

Solution

To increase the availability for the users, Statistics Sweden created the web application REGINA, enabling users to easily enter the database and to search for information about administrative geographies and their changes over time. REGINA is built on an open web based interface.

There are currently twelve different types of geographies available in REGINA, both administrative and statistical. For some geographies, the user can see all the changes from 1952 onward. The search can be specified to one specific type of geography, type of change or period. One typical query could be which municipalities have ceased to exist since 1970.

Since 2017, there is a view service connected to REGINA, where the user can display the geographies on a map, sorted by type of geography and reference years. The view service currently contains 12 types of geographies.

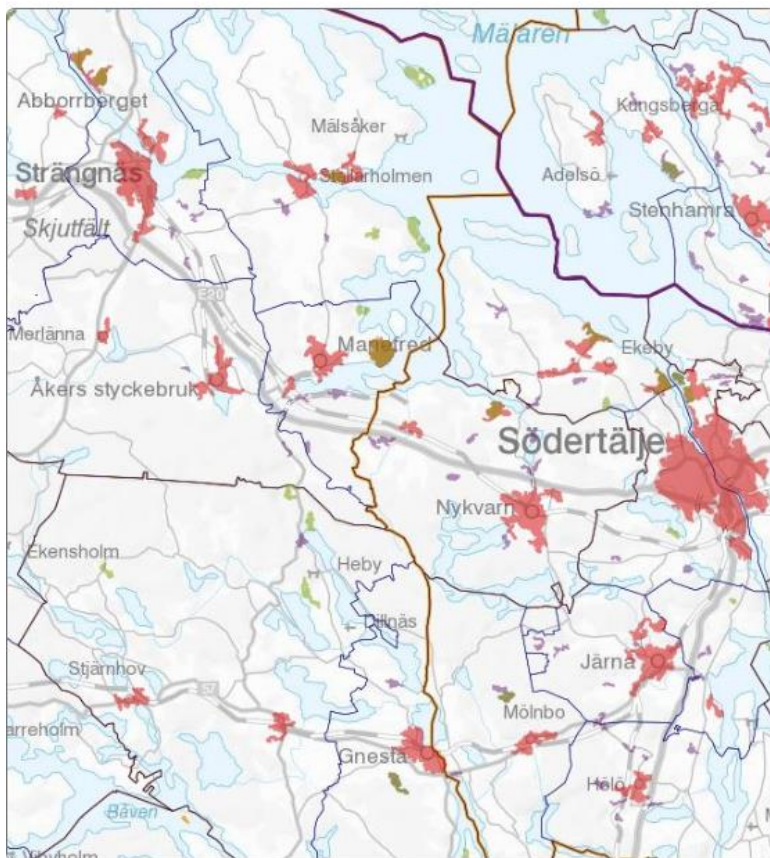


Figure 1: REGINA web service with regional divisions

Län Kommun Församling

Länskod Länsnamn

Bakåt Framåt 01 Stockholm Sök

Hjälp

1952 01 Stockholm stad

16 - Utvidgas och byter namn

1968 01 Stockholms län
01 Stockholm stad utvidgas och byter namn till 01 Stockholms län.
Hela 02 Stockholms län upphör till 01 Stockholms län.

15 - Utvidgas

1971 01 Stockholms län
03 Uppsala län ger del till 01 Stockholms län.
04 Södermanlands län ger del till 01 Stockholms län.

14 - Ger del till

1971 03 Uppsala län
01 Stockholms län ger del till 03 Uppsala län.

Figure 2: REGINA database interface showing the history of the county of Stockholm since 1952. Each change is recorded and described as an event in the database.

The Excel and PDF files, which were published before REGINA was launched, are still being updated and published at the website, for those users that prefer them.

Result

REGINA provides a better service to users by increasing access to the data and providing a better representation, for all sorts of users. Through the map service, accurate and detailed boundaries of the regional division can now be accessed as a complement to the lists.

More information

Web sites, only in Swedish:

- REGINA web service: <http://regina.scb.se/>
- Statistics Sweden's view service with regional divisions: <http://geodata.scb.se/reginawebmap/main/webapp/>

Contact information

Karin Hedeklint, Statistics Sweden, mark.vatten.gis@scb.se or scb@scb.se

(C3.2) Open geography portal by ONS (United Kingdom)

Keywords: Principle 3, management of statistical and administrative geographies, open data, linked open data

This use case mainly refers to requirement 3.1 to set up and maintain a consistent framework of national statistical and administrative geographies. Particularly this case demonstrates practical implementation of recommendation 3.1.2, that all national administrative, statistical and functional geographies with relevance for production and dissemination of official statistics should be provided as authoritative geospatial data in compliance with the technical specifications of INSPIRE. The case is also relevant for requirement 4.5, exploring the potential of Linked Open Data.

Note: This case was not prepared for or by the GEOSTAT 3 project but is considered as good practice by the project group.

Introduction

Open geography is a framework that provides the definitive source of geographic products, tools and services to a range of customers, including statistical producers and the private sector. These products can be accessed via the Open Geography portal (compliant with the EU INSPIRE directive to deliver harmonised spatial data across the EU) or with additional information via the categories in the main products page.

The ONS Geography Linked Data site is the access point for information on statistical geographies required to support the use of official statistics. It is designed to allow users to discover, view and use geospatial data. The site is complementary to the Open Geography Portal. It allows access directly to data within the geography products, in machine-readable form and using an Application Programming Interface.

More information

<https://www.ons.gov.uk/methodology/geography/geographicalproducts/opengeography>

Contact information

ons.geography@ons.gov.uk

(C3.3) European Location Services (ELS) – central access to harmonised, pan-European, authoritative geospatial information and services (EuroGeographics)

Keywords: Principle 3, pan-European service, statistical and administrative geographies

This use case mainly refers to the recommendations provided within Principle 3, Common geographies for production and dissemination of statistics. Mainly the recommendations provided under the requirement 3.1, set up and maintain a consistent framework of national statistical and administrative geographies and requirement 3.2, Improve maintenance of the European framework of statistical geographies – could be highlighted. Particularly, recommendation 3.2.2, that all European geospatial agencies are encouraged to support the current work on developing Open European Location Services (Open ELS) coordinated by EuroGeographics – is to be addressed.

The implementation of Principle 1 is a condition for the full implementation of principle 3 to allow for the flexible aggregation into any output geography. Particularly, all recommendations under requirement 1.1 – Use data from National Spatial Data Infrastructures – are a prerequisite for harmonised, pan-European, authoritative geospatial information and services.

Introduction

National conditions regarding the quality and the accessibility of European geospatial data vary from member to member, creating challenges for pan-European projects. INSPIRE aims to compile a European Union spatial data infrastructure for the purposes of EU environmental policies and policies/activities that may have an impact on the environment. However, having also a single access point for authoritative geospatial information, harmonized technical and business aspects, and licensing and pricing conditions of pan-European geospatial products and services, would allow an easier and faster implementation of solution strategies.

A core strategic goal of EuroGeographics is to facilitate access to and the use of its members' geospatial data and services. The vision for developing European Location Services (ELS) is to provide an access point for harmonised information and services from European National Mapping and Cadastre Authorities (NMCAs). Based on results of the European Location Framework (ELF) project that is already concluded, EuroGeographics aims to develop operational European Location Services. A first step is to develop operational Open ELS services under the CEF funded Open ELS project (project duration from May 2017 – April 2019; www.openels.eu).

Description of problem

At this point, no comparable European services exist for providing authoritative and harmonized pan-European data and services, by a single access point. Nevertheless, European NMCAs believe in the potential of European Location Services (ELS), and that there is an inherent demand for these.

Currently there is the need to approach multiple countries, negotiate multiple licenses or pay multiple fees. Regarding the EU legislation, the most significant pieces are the Public Sector Information (PSI) Directive, the INSPIRE Directive and associated technical rules, and key pieces of Copyright and Database Right Directives.

Whilst there are strong indicators of interest, there is no proven market willing to pay high fees for such authoritative and harmonized pan-European data and services. Therefore, centralized funds will be required to get the services' platform up and running. A business model for sustaining such funds may have to incorporate the potential to lower overall costs.

As a practical example of INSPIRE implementation, the ELF Project has supported the delivery of prototyped national web feature services and provided valuable feedback on the data specifications as they are implemented in different countries. It has delivered a technical infrastructure to incorporate data content into an application environment, as well as tools for harmonization, edge-matching and identifying areas of interest and products. It has also delivered some test services for specific use cases.

Solution

The planned European Location Services (ELS) shall cover the whole of geographic Europe and would offer much more than INSPIRE compliant datasets. ELS data would comply with consistent selection criteria for different scales, comprehensive quality requirements and edge matching on European borders.

ELS shall become a gateway to pan-European harmonised maps, geographic and land information from national sources. As a single source of official, quality-assured data from Europe's NMCAs and single point of access for licensing official data from the agencies of different countries, it shall provide harmonised data quality, specifications and standards, as well as a harmonised pricing and licensing. Furthermore, the spatial data will be INSPIRE-compliant and harmonised at a cross-border and pan-European level. The information and services based on user requirements enable the discovery, view, download and use of geospatial information, and make data integration possible.

The transition of ELF into the operational ELS is still work in progress.

Open ELS is a two-year CEF funded project that started in May 2017. The product and service development in the project is a mix of customer oriented and "supply driven" (available open data) approach. It focusses on the use of authoritative geospatial information by providing certainty about what is free, what is charged for and under what terms, conditions, and package these services are being developed to new open cross-border product and services. Furthermore, it will respect national policy, legislative and business requirements. The benefits will be open geospatial information from official national sources that are easy to find, accessed and re-used.

The current **prototyped products of ELS** are:

- 1) ELS View Services (ELS Topographic Base Map, ELS Cadastral Index Map, Web Map Services on the INSPIRE-compliant national data),
- 2) ELS Download Services (Web Feature Services providing access to the INSPIRE-compliant national data by a cascading architecture); data themes: Administrative Units, Geographical, Names, Buildings, Transport Networks, Hydrography, Cadastral Parcels, and Addresses
- 3) ELS GeoLocator (georeferencing service with Gazetteer Extension)

Result

If successfully implemented the operational European Location Services (ELS) will save time and costs. There will be no need for further harmonization, it will encourage reuse of public sector information (PSI) and there will be the option to consume the information as a services' platform that simplifies information management.

EuroGeographics recognizes that it may take a few years to accomplish this programme. It requires support and financial investment to co-ordinate and enable the development of ELS. The next steps will be defined in the first half year of 2019.

European and International institutions require pan-European harmonized geospatial information to support policy development and decision-making. Such decision-making depends on data provided by acknowledged official sources. Therefore, ELS could support these needs by providing harmonised pan-European authoritative geospatial information and services for this purpose.

More information

- 1) <https://eurogeographics.org/products-and-services/european-location-services/>
- 2) <http://openels.eu>
- 3) <http://locationframework.eu>

Contact information

EuroGeographics: contact@eurogeographics.org

(C 3.4) Exploring OGC Discrete Global Grid Systems DGGS (EFGS)

Keywords: Principle 3, Discrete Global Grid Systems, OGC standards, interoperability, statistical output geographies

This use case mainly refers to recommendation 3.3.3 provided within Principle 3, stating that the geospatial and statistical communities should monitor the development of the DGGS and its application closely in order to prepare for a possible future implementation of the grid system for national and European data.

Introduction

An Abstract Specification on Discrete Global Grid Systems (DGGS) were published by the OGC (Open Geospatial Consortium) in August 2017. The abstract does not specify a single set of grids but specifies 18 requirements that spans a universe of valid global grid systems. The requirements can be divided into two main topics; 1) the reference frame elements and 2) the functional algorithms. They again are divided into the following chapters:

6.2	DGGS Reference Frame Elements	6.2.1	Global Domain
		6.2.2	Tessellation Sequence
		6.2.3	Area Preservation
		6.2.4	Cell Structure
		6.2.5	Tessellations
		6.2.6	Spatial Referencing
6.3	DGGS Functional Algorithms	6.3.1	Quantization Operations
		6.3.2	Algebraic Operations
		6.3.3	Interoperability

Figure 1: Overall chapters and subchapters for describing a Discrete Global Grid System.

In front of the EFGS 2018 conference, the Global Forum for Geography and Statistics¹ challenged all participants to make population statistics on hexagonal grid cells (ISEA3H16).

The ISEA3H16 hex grid was produced using the free and open source software DGGRID, by kind help and support of Prof. Kevin Sahr at Southern Oregon University Computer Science Center.

Results

Results from the hexgrid challenge were presented as part of the GFGS voluntary project named Project FairHair (See appendix). In summary, results show that 16 countries and territories were able to make population grid statistics on hexagonal grids within a short time, with limited guidance and restrictions.

Once hexgrids had been uploaded to a spatial database, it was quite easy and fast to execute a few lines in SQL and have the results presented in standardized population groups. Most participants had used the classification legend developed in the Geostat 1A project.

¹ GFGS task force: <https://www.efgs.info/about-efgs/global-forum-for-geography-and-statistics/>

Hexgrids by country were made by using borders from GADM (Global Administrative Boundary Database). The borders and coastline in GADM turned out to lack some populated islands. Country based hexgrids should hence cover all territories, also sea territories. These hexgrids should also include a buffer zone around outer boundaries, due to geographic inaccuracy of the boundaries. Alternatively, larger hexgrids should be intersected with more accurate national datasets on administrative boundaries.

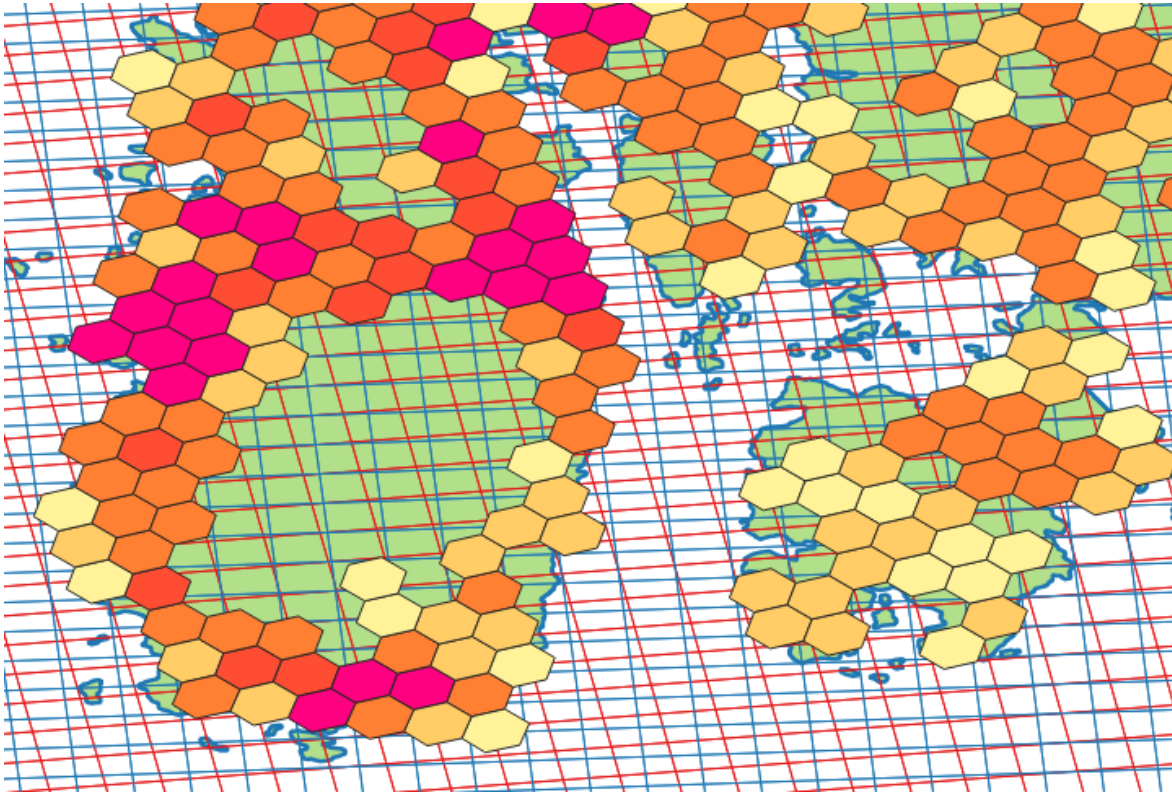


Figure 2: Example map (EPSG: 3226)

Avaldsnes area in Norway. ISEA3H16 hexagons with population counts (EPSG:4326 WGS84), 1x1 km INSPIRE grid (blue squares, EPSG: 3035 - ETRS/LAEA Europe) and national 1x1 km grid (red squares, EPSG:32633 WGS84/UTM33N)

Immediate reactions to hexgrids are twofold. On one side, you have square grids systems, with their historical background and hierarchy that are easy for humans to grasp. Square grids furthermore have a history of grid statistics. However, it is not possible to have a DGGS based upon square grid cells, without breaking some of the OGS requirements.

Hexgrids on the other hand are statistically sounder, in that the centroid of each cell has an equal distance to the centroids (representation points) of the neighboring polygons. Hexgrids are also said to be 40 per cent more efficient in storing and capturing information. However, standardized tools for operating between different resolutions or between different polygons must be developed.

Further work

Although the OGC DGGs Abstract Specification describes valid grid systems, it is not defining how this should be done. There exists a lot of software and script (e.g. DGGRID, OpenEAGGR, HEALPix, dggridR, SCENZ-Grid, PYXIS, QGIS plugin, Proj.4, Sphrekit, SCRIP) that fulfil parts of the DGGs requirements described by OGC. However, neither are there any institute for certifying software or any mechanism that do evaluation of such.

Some countries have a history of grid statistics attached to national square grids. It is not likely that these countries will suddenly start using a grid according to the OGC DGGs requirements. The preliminary tests and comparisons of different ways of reporting statistics to an OGC DGGs compliant grid should continue.

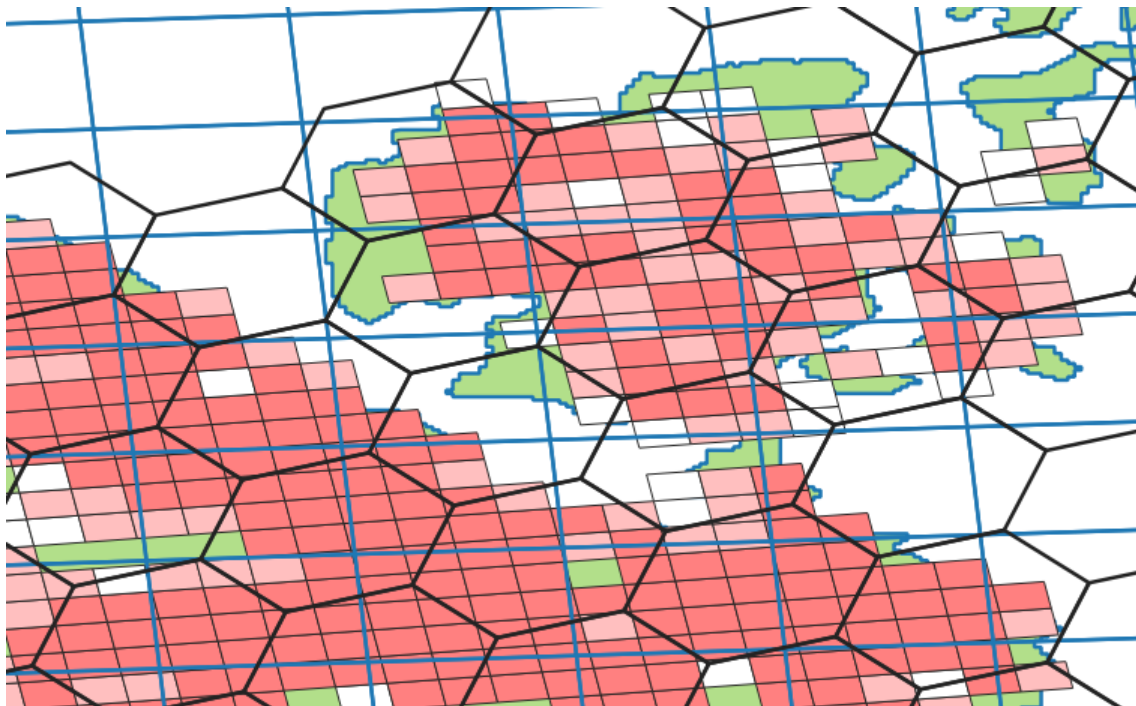


Figure 3: Example map (EPSG: 3226). Stavanger area in Norway

ISEA3H16 hexagons (EPSG:4326 WGS84), 1x1 km INSPIRE grid (blue squares, EPSG: 3035 - ETRS/LAEA Europe) and national 250 x 250 m grid with population counts (EPSG:32633 WGS84/UTM33N)

The GFGS aims at testing and developing DGGs in 2019 and onwards. A paper on comparison of hexgrids, as an example of part of a DGGs, and square grids will be made and presented.

More information

OGC DGGs Abstract Specification : <http://docs.opengeospatial.org/as/15-104r5/15-104r5.html>

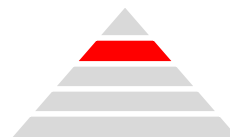
GFGS task force: <https://www.efgs.info/about-efgs/global-forum-for-geography-and-statistics/>

DGGRID : <http://www.discreteglobalgrids.org/software/>

Contact information

Vilni Verner Holst Bloch, Statistics Norway, vilni.verner.holst.bloch@ssb.no or ssb@ssb.no

4 Principle 4: Statistical and geospatial interoperability – Data, Standards and Processes



(C 4.1) Geospatial Reference Architecture – a tool for managed utilisation of geospatial information (Finland)

Keywords: Principle 4, data integration, information architecture, logical data warehouses, GSBPM

This use case, describing the Finnish Geospatial Reference Architecture, refers to all principles. Its implementation will fulfil the requirements of Principle 4 (and its requirements and recommendations). It relies completely on well-managed implementation of Principles 1 to 3 and it is in line with their requirements and recommendations. Principle 5 is also included as high-level goals, but not yet as planned actions from the architecture point of view. The first version of the geospatial reference architecture needs to be further developed and observed from the GSGF perspective, but benefits should also be taken into account the other way round – the enterprise architecture method can offer a way to implement the GSGF in the NSIs.

Introduction

The reference architecture work defined the vision for the desired state and its actors for the processing of geospatial information utilised and produced in Statistics Finland's statistics production. Measures were documented in the reference architecture for attaining the desired state according to the vision. The key perspective is to improve interoperability within Statistics Finland and outside the agency in the operating environment of geospatial information.

Description of the problem

The processing of geospatial information utilised and produced in statistics production is implemented with overlapping statistics-specific solutions that may have led to mutually non-uniform results and overlapping work. There has been no comprehensive view of geospatial information linkages of the entire statistics production and this has made it difficult to utilise fully and particularly administer geospatial information of statistics.

The integration of statistics and geospatial information is placed nationally within the activity of several authorities. There is no clear view of the division of work between different authorities.

Solution

The first version has been produced of the geospatial reference architecture. It is a tool for controlled utilisation and production 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 production of geospatial information
- Geospatial information is defined in the logical geospatial data warehouse
- Geospatial information is used uniformly

- Geospatial information is used with a customer-oriented approach
- Statistics Finland cooperates with other producers of geospatial information.

The framework for the architecture work is the national public administration recommendation JHS 179 Planning and development of business architecture. It is also based on the reference architecture of public administration's geospatial information, which is a desired state description and development plan advancing the national interoperability and joint use of geospatial information

Results

Strategy map

The strategy map of Statistics Finland's geospatial reference architecture summarises the vision of geospatial information, the drivers behind the change, the strategic goals and the detailed level objectives.

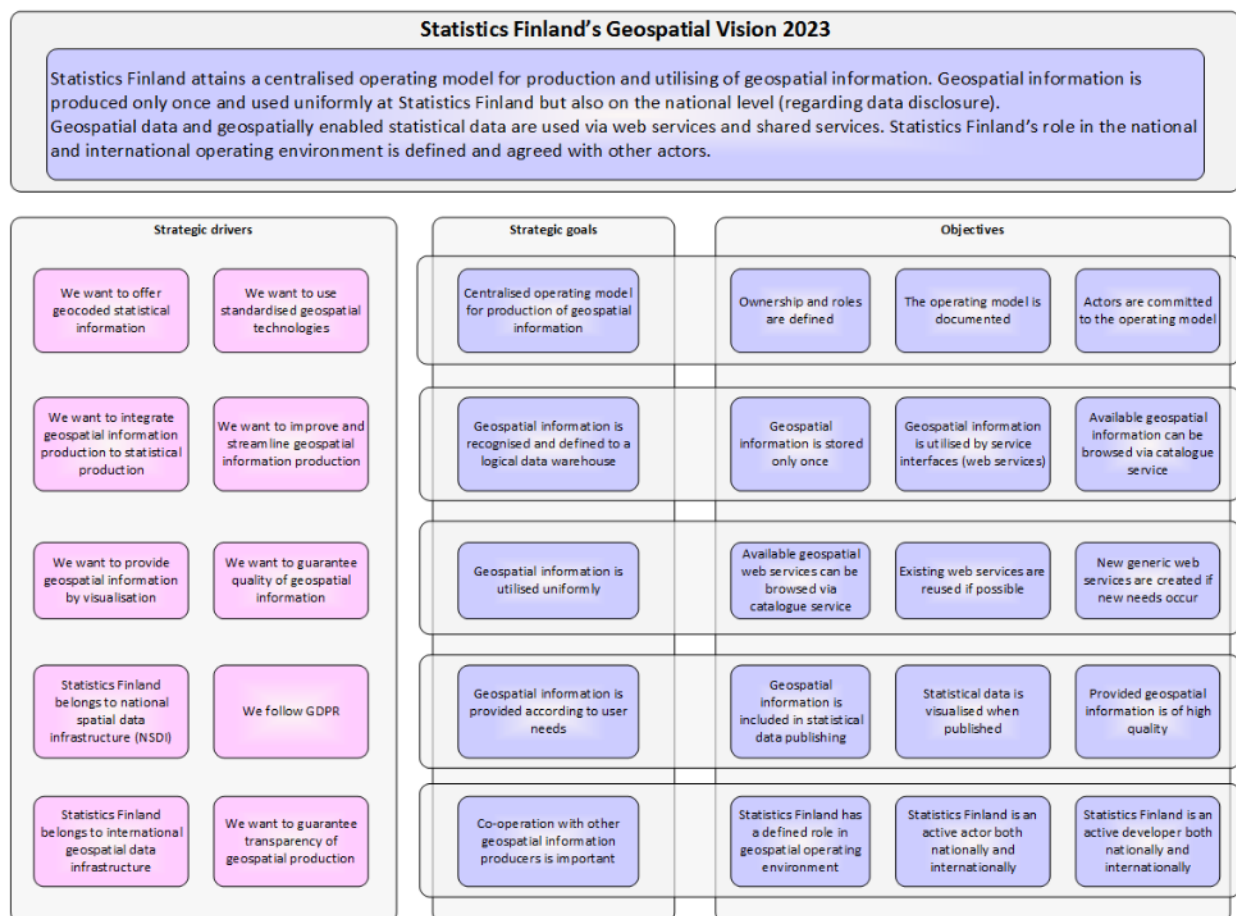


Figure 1: Geospatial Strategy Map of the Geospatial Reference Architecture of Statistics Finland

Strategic goal	Description
Centralised operating model for production of geospatial information	Ownerships, roles and actors of spatial information are defined. The activities related to geospatial information are centralised and the operating model is documented so that it is transparently available to all involved in the production and distribution of geospatial information. Geospatial information actors must be committed to the centralised operating model.
Geospatial information is defined in the logical geospatial data warehouse	Geospatial information and location information are stored in the centralised geospatial data warehouse. The processing of geospatial information is based on the joint concept model. Data of the logical geospatial data warehouse are used through interfaces (in the case of geospatial information through OGC interfaces).
Geospatial information is utilised uniformly	Geospatial information is used at Statistics Finland and as far as possible on the national level uniformly through generic geospatial information services. Generic services are produced for new needs. The desired state services of geospatial information are presented in the service map. Statistics Finland also utilises general use services produced by other geospatial information actors. Correspondingly, Statistics Finland produces such services that are available to other geospatial information actors.
Geospatial information is provided according to user needs	The aim is to provide high-quality geospatial information as part of the distribution of statistical data whenever location can be linked to statistical data. The data are provided through the services and they are available to customers in visual form, for example, through map interfaces. Customers can combine various statistical data on the basis of location information.
Cooperation with other geospatial information producers	Statistics Finland defines its role in the operating environment of geospatial information. Statistics Finland can utilise more outputs of other actors and focus on the development of the outputs according to its role. These outputs are correspondingly available to other geospatial information actors, for example, as generic services or more high-quality data. Statistics Finland is active nationally and internationally.

Stakeholders of geospatial information

Stakeholder	Description
International	<ul style="list-style-type: none"> • EU Eurostat, WG Integration of Statistical and Geospatial information • UN ECE High Level Group for Modernisation of Official Statistics HLG-MOS • United Nations Committee of Experts on Global Geospatial Information Management UN GGIM • United Nations Expert Group on the Integration of Statistical and Geospatial Information UN EG ISGI • European Forum for Geography and Statistics EFGS • GEOSTAT consortiums

National	On the national level, around 20 key actors were identified. The most significant of these is the National Land Survey of Finland for cooperation and as producer of data. National collaboration to advance integration of statistics and geospatial information must be strengthened. A shared national Finnish view in the above-mentioned international connections requires a support structure for national cooperation. We suggest the establishment of this kind of group.
Statistics Finland's internal	Over 40 different statistical systems, products or activities were identified at Statistics Finland where the processing of geospatial information is part of statistics production.

Conceptual model

The logical geospatial data repository is one of Statistics Finland's six logical data repositories in the processing stage, which is linked to other logical data repositories as concerns location information of statistical objects.

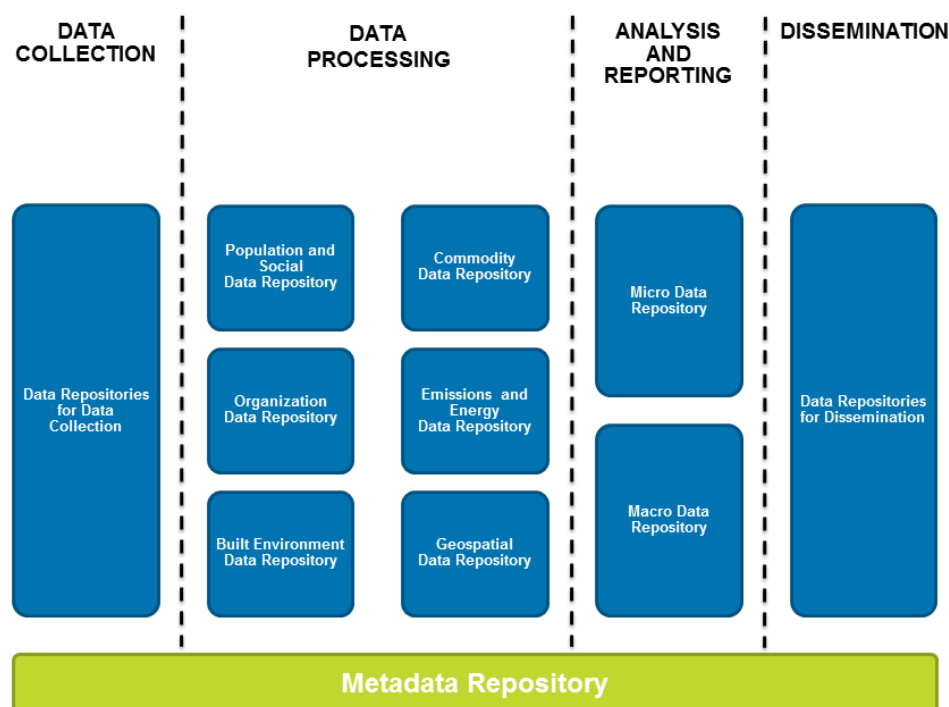


Figure 2: Statistics Finland's logical data warehouses

A less detailed level conceptual model for geospatial information was made in connection with the reference architecture work. Statistics Finland's geospatial information solutions must comply with the data model described below to attain interoperability of service and integrability of data.

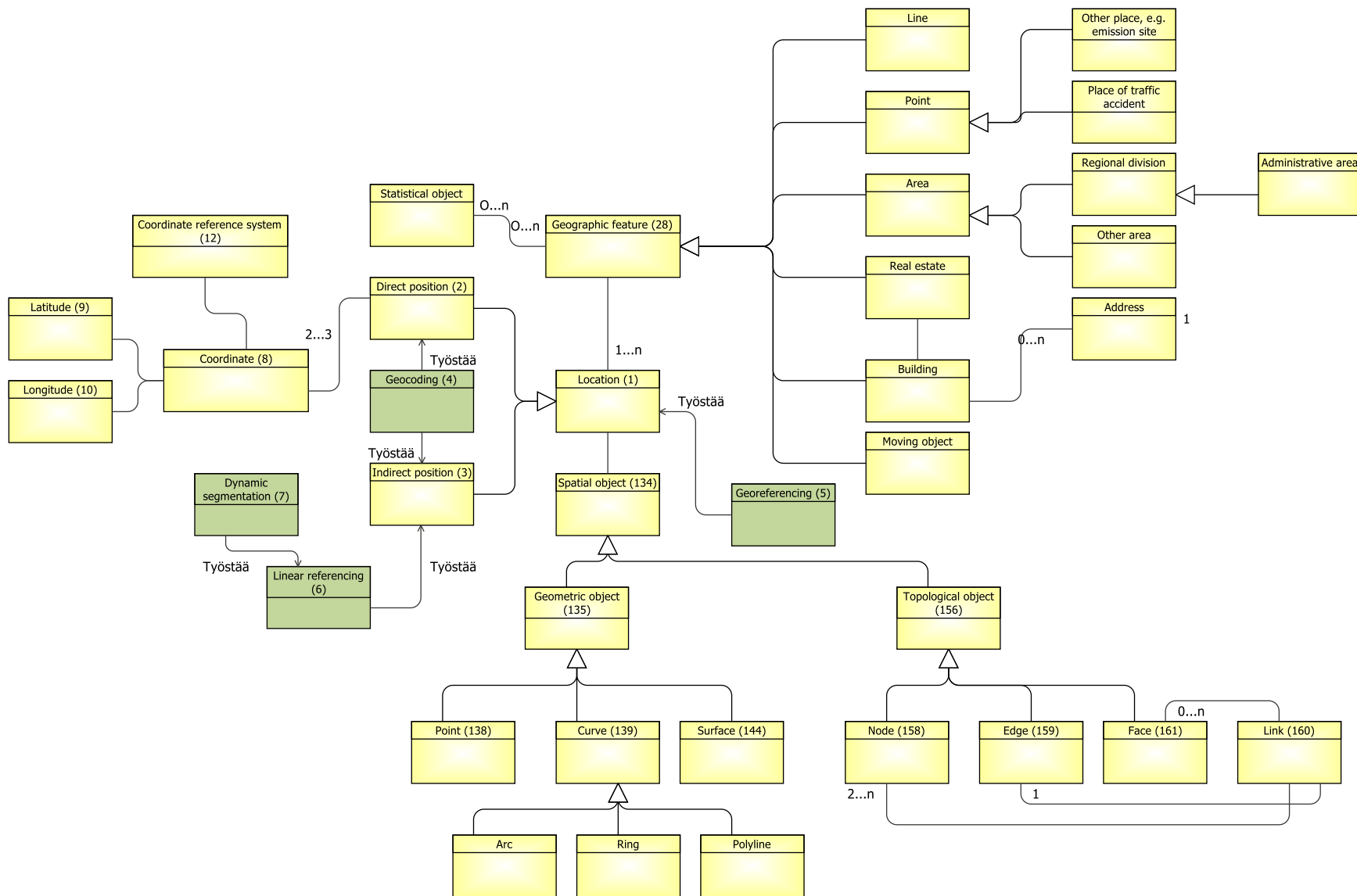


Figure 3: Conceptual model for spatial information (to be developed)

The conceptual model comprises statistics production concepts and geospatial information concepts specifying them. The *geographic feature* is at the centre of the conceptual model. It is a *statistical object* that has a *location*. The definition of location, or *georeferencing*, can consist of *direct position* or *indirect position*. If the indirect position of a geographic feature, e.g. address data, needs to be direct position data (coordinates), geocoding can be performed as the operation. The conceptual model is specified in connection with the implementation of the logical geospatial data warehouse, when a centralised data warehouse is built for geospatial information. The logical geospatial data warehouse is built so that it can be integrated with warehouses and services of other geospatial information actors.

Actors

Geospatial information actors are classified at Statistics Finland into three groups:

1. **Steering:** Statistics Finland's Management Group is responsible for the strategic objectives and the development portfolio group for the selection of development objects. The ownership of geospatial information must be named.
2. **Development, administration and maintenance:** A responsible party is named for geospatial information. The responsible party is a multi-professional team that removes silos inside the organisation and enables development of geospatial information over unit boundaries. The multi-professional team is responsible for different sub-areas of architecture (business architecture, data architecture, information system architecture and technology architecture). Understanding the theoretical competence of geospatial information is included in all roles of the responsible party.
3. **Statistics production:** Data collection, statistics production and data dissemination use geospatial information from the logical geospatial data warehouse. The data are used and distributed through geospatial information services maintained by the responsible party.

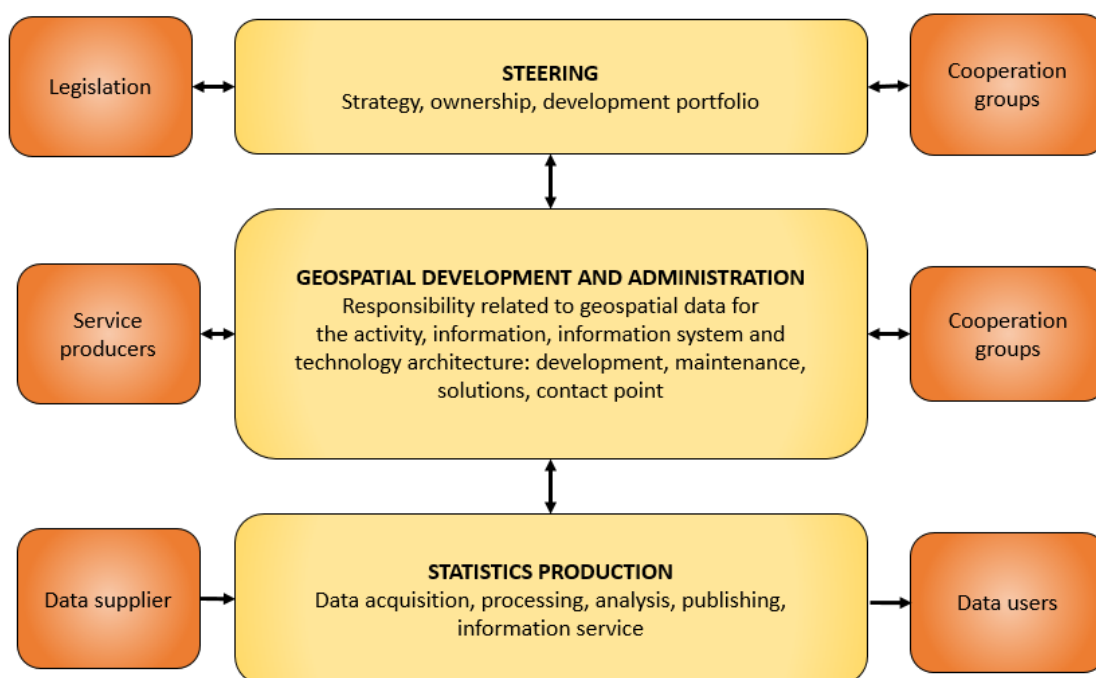


Figure 4: An illustration of geospatial information actors (on a coarse level)

Service map

Geospatial information services required by statistics production are classified in the service map according to the Generic Statistical Business process Model (GSBPM) phases. The services are generic services that can be extended as needed by statistics production or customers. The aim is to utilise services produced by other organisations in the implementation. New services are developed whenever necessary. Statistics Finland can produce part of the services so that they can also be utilised by external actors.

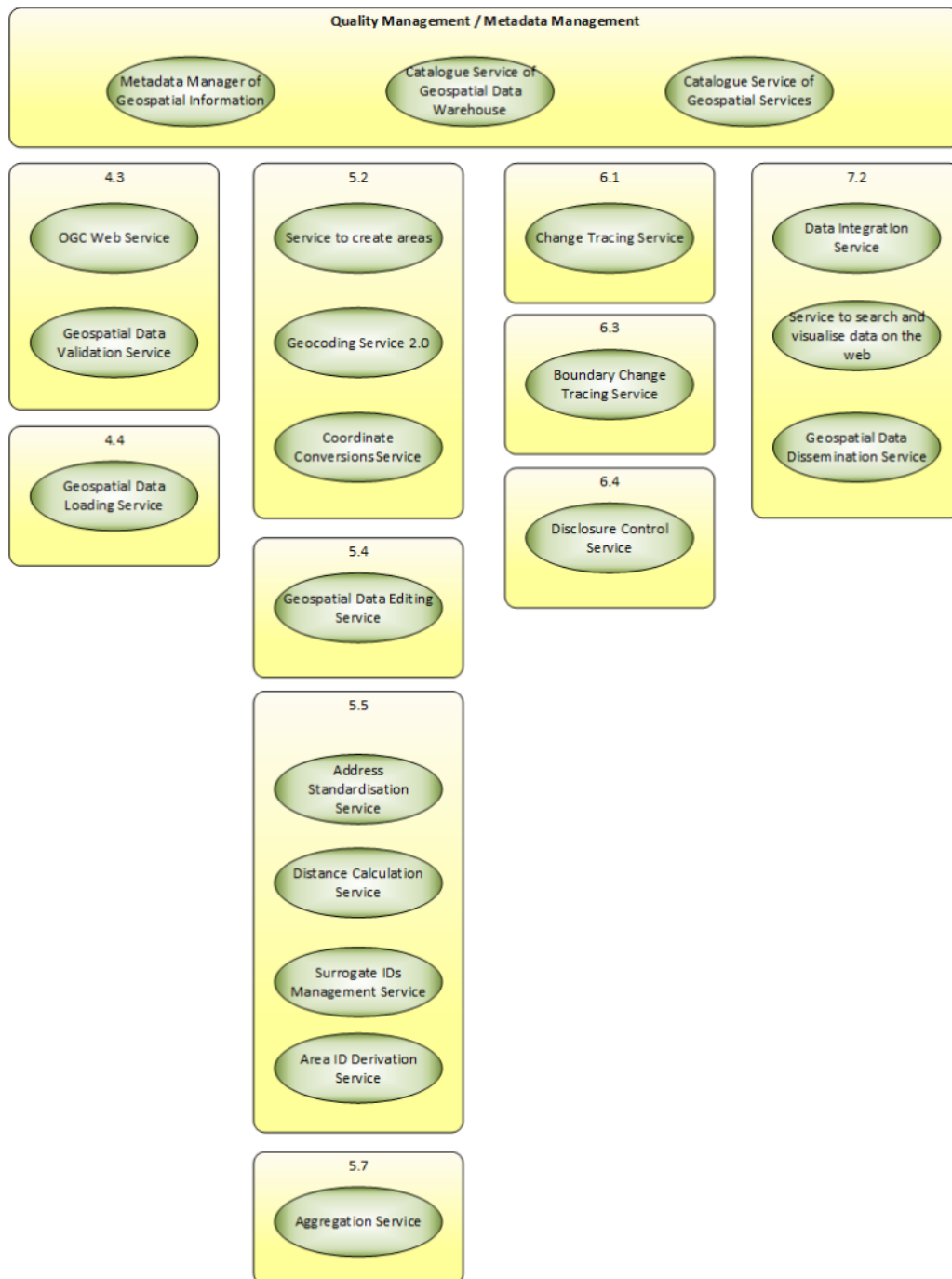


Figure 5: List of planned services according to the GSBPM phases

Road Map

In connection with the geospatial reference architecture work, more detailed level measures were listed for attaining the change sought with the reference architecture. The measures to be made in the first phase are compiled into development packages in the development road map. The first phase (during 2019) measure packages are

- Definition of geospatial information actors
- Temporary maintenance of geospatial data reference architecture and creation of the control model
- Implementation of the logical geospatial data warehouse and the first services
- Further development of the logical geospatial data warehouse and services
- Development of metadata for geospatial information
- Implementation of the map service

Contact information

Antti Santaharju, Statistics Finland, antti.santaharju@stat.fi or info@tilastokeskus.fi

Rina Tammisto, Statistics Finland, rina.tammisto@stat.fi or info@tilastokeskus.fi

(C 4.2) Making SDMX fit for INSPIRE – How statistical tools can deliver INSPIRE compliant data and metadata (Eurostat)

Keywords: Principle 4, data integration, INSPIRE compliance, interoperability, metadata

This use case covers, fully or in part, the following recommendations and requirements under Principle 4:

- *Requirement 4.1 including recommendations 4.1.3 and 4.1.6.*
- *Requirement 4.3 including recommendations 4.3.2 and 4.3.3.*
- *Requirement 4.4 including recommendation 4.4.1 and 4.4.4.*

Introduction

Geospatial and statistical data and metadata are shared using different data formats, exchange methods and dissemination standards. In Europe, geospatial information is shared using the spatial data infrastructure INSPIRE while statistical information is exchanged following Standard for Data and Metadata eXchange (SDMX). Defining a mapping between these two standards is essential to support the combination of these two types of information and maximise the re-use of existing and accepted data infrastructures for statistics. This would allow statistics organisations to enrich their datasets with geospatial information and re-use INSPIRE enhanced tools. The combined data creates greater value and INSPIRE services illustrate statistics in a visually enhanced manner.

Eurostat carried out a successful pilot study to integrate INSPIRE concepts into SDMX in the context of the Census 2021 data collection. A mapping between the INSPIRE themes Population distribution and Statistical Units on the one hand and Census data and metadata modelled in SDMX for the exchange of census information on the other was defined and will be implemented in Eurostat's SDMX enabled data exchange infrastructure, the Census Hub. As a result Statistical Offices will implement automatically the requirements from INSPIRE without disruption of their established production systems and without double data sharing burden.

Description of problem

The INSPIRE directive and its implementing regulations require public authorities in Member States holding spatial data to share these data via national spatial data infrastructures (NSDI). Population distribution and census information is one of the themes that is covered by INSPIRE and as a consequence Statistical Offices holding census information have to meet the legal requirements of INSPIRE. Specifically NSIs need to meet the following requirements:

- 1) Transformation of population data into an INSPIRE data model (data interoperability);
- 2) Provision of network services:
 - a) Discovery service to search for data and services by means of INSPIRE metadata;
 - b) View services to view the data;
 - c) Download service to obtain copies of the data;
- 3) Creation of INSPIRE metadata on population grid data and the above services;

The INSPIRE roadmap requires Member States to fully comply with INSPIRE by the end of 2021. This means that for the 2021 round of population and housing censuses in the EU, statistical offices will

have to share census data according to INSPIRE legislation in addition to the existing statistical dissemination infrastructure based on SDMX.

The goal for the EU wide 2021 census has been to minimise the effects of this double obligation on Member States and to maximise the usability of the census information for the statistical and geospatial community.

Solution

The similarities between the Census Hub data infrastructure and an INSPIRE spatial data infrastructure are obvious. Therefore the potential of the Census hub to meet these requirements and to avoid duplication of infrastructures has been assessed, with positive results.

The conclusion of the analysis is that the Census Hub could be made INSPIRE compliant with fairly limited effort by using it for the transmission of population grid data and metadata from the Member States to Eurostat in the SDMX data model. The actual implementation of INSPIRE would only happen at the central level in Eurostat (see illustration).

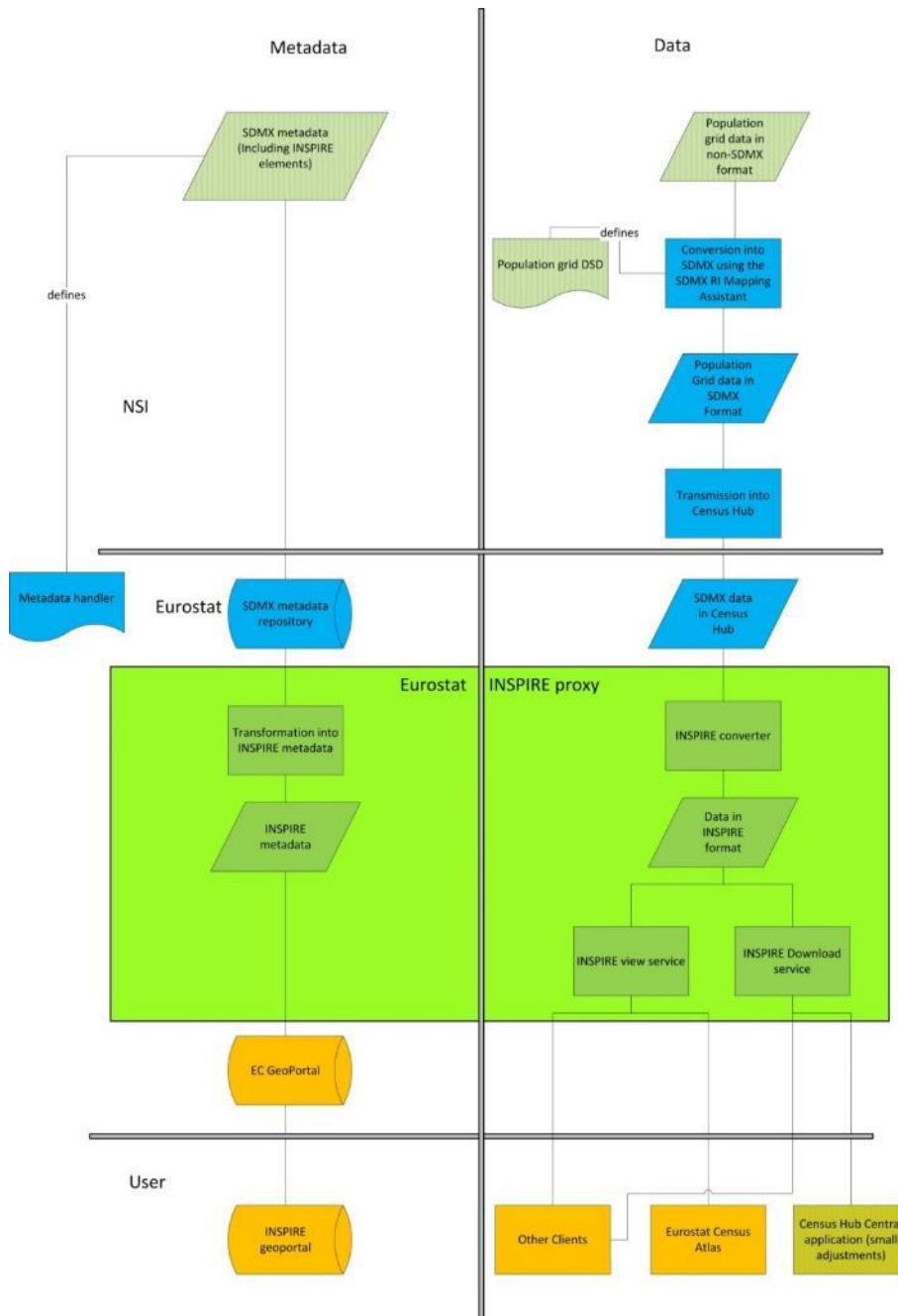


Figure 1: Architecture and workflow for INSPIRE compliant population grids from NSIs to Eurostat.

The orange components are existing INSPIRE components; the blue components are existing Census Hub and ESS Metadata Handler components. The green components need to be developed/ created for INSPIRE compliant population grid data. Hatched components need to be developed for SDMX transfer of population grid data, in parts independently of INSPIRE.

Workflow of data and metadata

The following workflow (see Figure) provides a more detailed description of what happens where and who does what, starting from the transmission of population grid data from Member States into the Census Hub until the sharing of INSPIRE data and metadata via INSPIRE services to the final users.

- 1) Member States create population grid data in their preferred format;
- 2) Member States, using the Population Grid DSD defined by Eurostat, and the data mapping tool of the SDMX RI, transform the data into SDMX format;
- 3) Member States, using the ESS Metadata Handler provide reference metadata on Population Grids including the additional INSPIRE elements;
- 4) The Census Hub pulls the population grid SDMX data into a central repository;
- 5) A converter as part of the INSPIRE proxy developed by Eurostat converts SDMX metadata into INSPIRE metadata;
- 6) A converter as part of the INSPIRE proxy developed by Eurostat converts SDMX data into INSPIRE data;
- 7) Eurostat sets up an INSPIRE download service using a local copy of the INSPIRE data. One of the clients of this download service will be the Census Hub central application;
- 8) Eurostat sets up an INSPIRE view service using a local copy of the INSPIRE data. One of the clients of this view service could be a Census Atlas developed by Eurostat.
- 9) INSPIRE metadata on data and services will be included into the INSPIRE geoportal via the European Commission INSPIRE discovery service.

Result

Eurostat has worked on the mapping of INSPIRE and SDMX concepts for data and metadata and has built Data Structure Definitions (DSD) and Metadata Structure Definitions (MSD) prototypes and examples.

Mapping INSPIRE metadata to the census ESMS

There is generally good correspondence between discovery metadata (INSPIRE) and reference metadata (SDMX). Many of the INSPIRE elements are derived from Dublin Core and in essence provide basic information on the title, the history and the content of the resource. As such, these elements are very similar to ESMS metadata elements with a similar purpose.

However conceptual overlaps alone are not sufficient for a consistent mapping of SDMX and INSPIRE. The following cases can be distinguished:

- 1) The metadata element is the same in terms of semantics and syntax, and a 1:1 mapping can be made;
- 2) The metadata element cannot be mapped 1:1. This can have several reasons:
- 3) The semantics is different, e.g. the scope of an element is narrower or wider in one of the standards;
- 4) The encoding of the element is different (e.g. free text vs code lists).
- 5) The information from INSPIRE is entirely missing in the ESMS and needs to be added.

A few INSPIRE elements mainly describing the resource could be directly mapped to census MSD elements. To avoid complex partial mappings due to undercoverage, overcoverage or different syntax, the remaining elements were simply added as additional INSPIRE concepts to the Census MSD. This approach proved to be very successful and a full coverage of all INSPIRE elements in the new, extended Census MSD could be reached.

It should be noted that the mapping of INSPIRE metadata elements to the Census MSD can be easily extended to other statistical domains, as INSPIRE metadata elements are not topic specific. As a result, the extended MSD is not restricted to Census data.

One important aspect is that all additional INSPIRE elements that are not yet covered by already existing, equivalent census MSD elements can be filled automatically using existing information (e.g. the spatial extent). As a result, no extra manual work on the side of NSIs related to INSPIRE metadata is expected.

ESS Metadata Handler

After the extension of the current Census Hub MSD with INSPIRE metadata elements, the transmission of national reference metadata will be supported by the European Statistical System Metadata Handler (ESS-MH) including those additional INSPIRE elements. The ESS-MH is an IT application that allows users (Eurostat and ESS) to produce and disseminate reference metadata files.

This means that NSIs will be able to transmit INSPIRE compliant metadata to Eurostat via the ESS-MH without the need to use INSPIRE encoding or understand INSPIRE metadata.

It will be Eurostat's responsibility to extract the INSPIRE relevant metadata elements from national reference metadata and consolidate and convert them into an INSPIRE compliant metadata file.

Mapping INSPIRE data models to the Census DSD

The mapping between the INSPIRE and SDMX data models for census statistics has followed a similar approach as for metadata. From the outset SDMX has been an important input for the design of the INSPIRE population distribution and demography data model, and as a result concepts and code lists are aligned to a certain extent. Nevertheless INSPIRE requires additional attributes and the actual naming of attributes is different due to generic INSPIRE requirements.

While at the conceptual level, SDMX and the INSPIRE annex on population distribution and demography are largely aligned, the actual data structures are different mainly due to the structural rigour of SDMX. This proved to be complex but could be solved successfully and there is now an extended Census DSD that accommodates the requirements of population grids under INSPIRE.

Contrary to the metadata, the data model of INSPIRE for census information is theme specific. A transfer of this specific model to other types of statistics is therefore not possible without further analysis.

In conclusion, conceptually the adaption of SDMX for INSPIRE implementation using Census Hub tools is completed. The next step consists in enhancing the existing tools e.g. for an INSPIRE download service, and create transformation tools from SDMX to INSPIRE. Once the census data are available in the Census Hub Eurostat will also need to create the actual metadata file and view services.

More information

<http://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census>

Contact information

Ekkehard Petri, Eurostat, ESTAT-GISCO@ec.europa.eu

(C 4.3) Publish data once and leave data at its source (Netherlands)

Keywords: Principle 4, table joining services, OGC standards, data integration, INSPIRE compliance, interoperability

This use case mainly refers to a number of requirements provided within Principle 4. Particularly it addresses the requirement to publish data once and leave it at its source to be reused many times, including use of SDMX and service-oriented dissemination through APIs to provide machine-readable open data. It also demonstrates the use of Table joining services as a means of merging geography and statistics.

Introduction

This guiding principle of *publish data once and leave data at its source*, means that all data, both geospatial and statistical, should ideally be published only once and published separate of each other. Since constructing the geometries of regions and compiling tables follow different pathways throughout the NSI. Thus implying that in the case of changing tabular or geometry data the construction of a combined dataset is automatically following the most recent changes.

Building tabular data and geometry in open data format allows machine-to-machine based transformation and integration of data. This idea of a single entry for a given dataset is already one of the pillars of INSPIRE but could also be applied in a broader scope to include also statistical information.

Implementing this guiding principle can help mitigate problems arising from duplication of data with unclear origin and actuality. It can also save time and resources through simplified publishing procedures. However, this will require increased use of tooling and services that automatically can merge tabular statistical data with geospatial data on administrative and statistical geographies online. Ideally, tailor made data fusion products can be created on-demand by the end user.

In a long-term perspective, endorsing this guiding principle could lead to a more efficient, service-based exchange of data between institutions on a national level and between the national and European level. Most NSIs have since long established platforms to store and disseminate official statistics (E.g. Statistical databases). Some of these platforms are built on common frameworks such as PX web. The rapid development of web protocols for data exchange and APIs have provided new opportunities to search and harvest data from these platforms in new ways, with greater flexibility and possible integration of statistical data in third party applications without a need to create physical copies of data.

Statistics Netherlands has its own client side application that joins the geometry to the tables. Secondly, Statistics Netherlands is working on implementing a Table joining Service that does the join on the server site with OGC services as output.

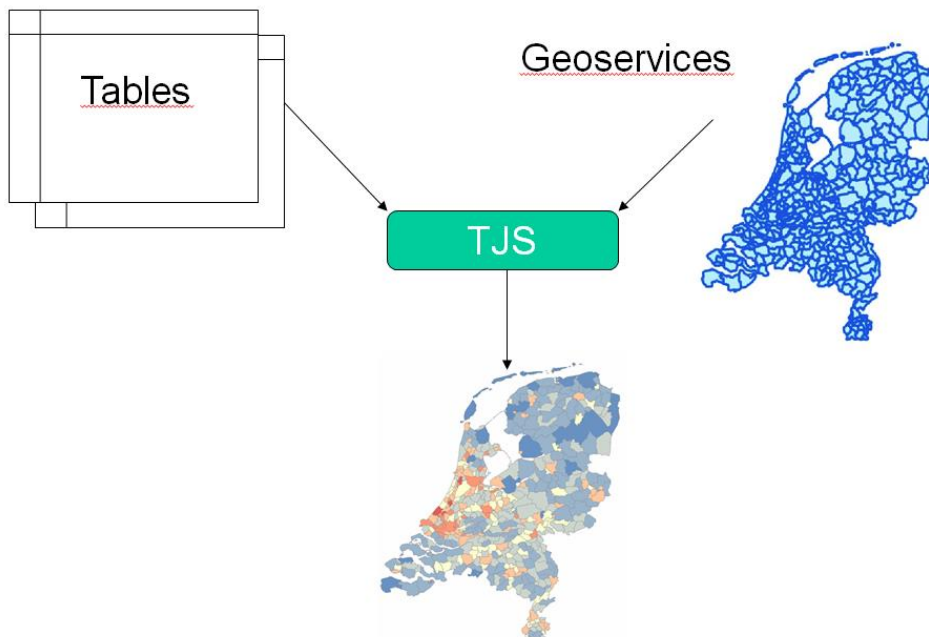


Figure 1: Principle of Table Joining Services TJS

Description of problem

If data is not left at its source, it means copies are made. In the case of geometry and statistics this often happens when GIS specialist link them together and publish the joined data a geoservice.

The problem arises when either of the datasets sources is being updated. In most cases, the geoservices is not updated after such an update, which leads to two published data sets that are not in line.

In most cases, the statistical tabular data is updated more frequently than the geoservice, because it is too cumbersome to again join the new table to the geometry and create a new geoservice. For instance at Statistics Netherlands, we only update geoservices of neighbourhood statistics once a year, whereas the source tables are updated 3 to 4 times a year.

Solution

The solution is to publish the separate datasets as separate machine-readable services in an open data format for tabular and geometry data. Special tooling will join this data automatically.

Tabular data

In the case of Statistics Netherlands, the tabular data is published according to the Open Data Protocol (Odata):

<https://www.cbs.nl/en-gb/our-services/open-data>

Data published according to this protocol is machine-readable. For instance, you can directly import the data into Excel when you have an Internet connection. However, there is also a so-called RESTful API, which is an interface for application developers that want to access the data in there programs.

Geometry

The geometry from the Dutch statistical units are published as OGC web services via the Dutch hosting organisation <https://www.PDOK.nl>. They are published in their original format (Asis) and in a INSPIRE harmonized format.

Joining the data automatically

At this moment, there are two examples of joining the data automatically by using the above-mentioned services. One is the example of a client side application that is directly connected to the StatLine application of tabular data and to the OGC web service of the geometry:

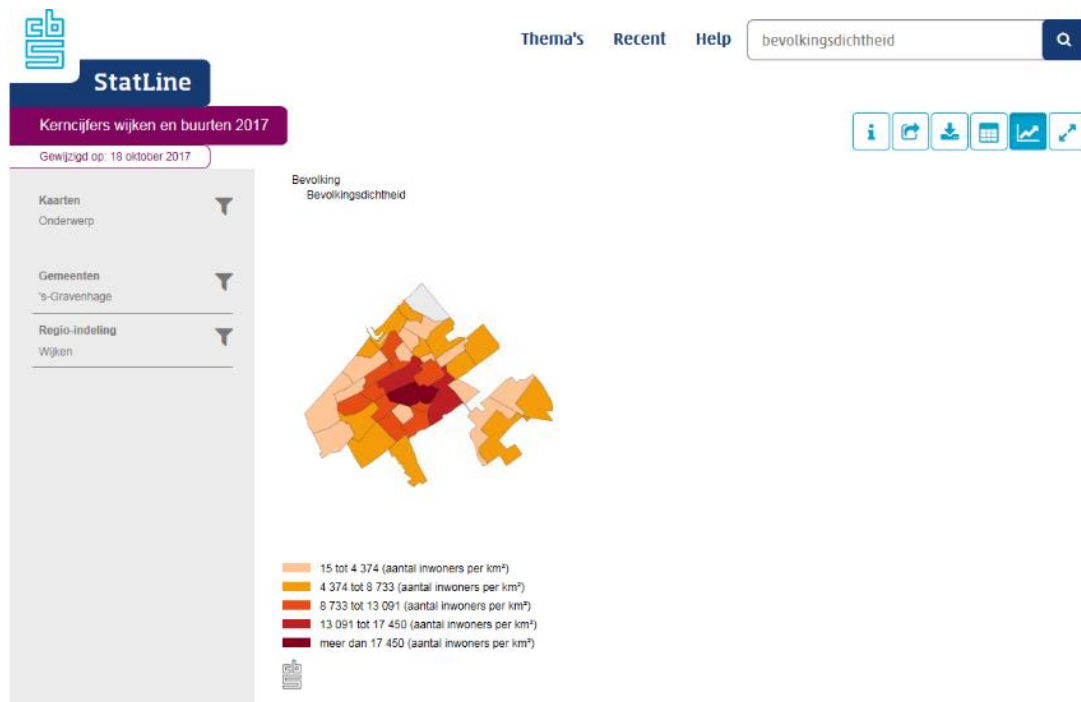


Figure 2: StatLine application

This example joins the population density in the table to the geometry of the neighbourhoods from the community of The Hague. The result is just an image presented in the online Statline application.

New development

The other example is the Pilot for a Table Joining Service. This service joins the tabular services and geoservices with a new OGC geoservice as output. This enables GIS users and application builders to directly use the joined data in their (GIS)-applications.

Result

The results have been described partly in paragraph 5. At this moment Statistics Netherlands is implementing a Table Joining Service at the hosting organisation for their geoservices <https://www.pdok.nl/>. The result is expected in the 3rd quarter of 2019.

More information

Odata protocol

<https://www.cbs.nl/en-gb/our-services/open-data>

<http://www.odata.org/>

<https://en.wikipedia.org/wiki/> Open Data Protocol

Metadata for the geometry of statistical units in original format (in Dutch):

<http://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/effe1ab0-073d-437c-af13-df5c5e07d6cd>

View and download via:

<https://geodata.nationaalgeoregister.nl/cbsgebiedsindelingen/wms?>

<https://geodata.nationaalgeoregister.nl/cbsgebiedsindelingen/wfs?>

Metadata for the INSPIRE harmonized of statistical units (in English):

<http://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/10d1153e-778f-4995-9b6c-7c69b196cccb>

View and download via:

<https://geodata.nationaalgeoregister.nl/su-vector/wms?>

<https://geodata.nationaalgeoregister.nl/su-vector/wfs?>

Example of client side application joining the tabular data and geometry:

<https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83765NED/map?ts=1513774593329>

Links for the result of the Table Joining Service Pilot:

<https://themes.jrc.ec.europa.eu/file/view/113290/impact-analysis-of-a-table-joining-service>

Contact information

Pieter Bresters, Statistics Netherlands p.bresters@cbs.nl or info@cbs.nl

(C 4.4) Linked Data based model of joint production process of statistical units (Finland)

Keywords: Principle 4, linked data, data integration, interoperability

This use case demonstrates interoperability as requested by Principle 4. It comprises examples of organisational interoperability, semantic and conceptual interoperability and also technical interoperability. It actually implements all the requirements of Principle 4 (and their recommendations), but especially requirement 4.5. The solution described requires well-defined and implemented requirement 3.1. It can be a basis for implementation of principle 5, especially requirements 5.2 (and recommendation 5.2.2) and requirement 5.4.

Introduction

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 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:

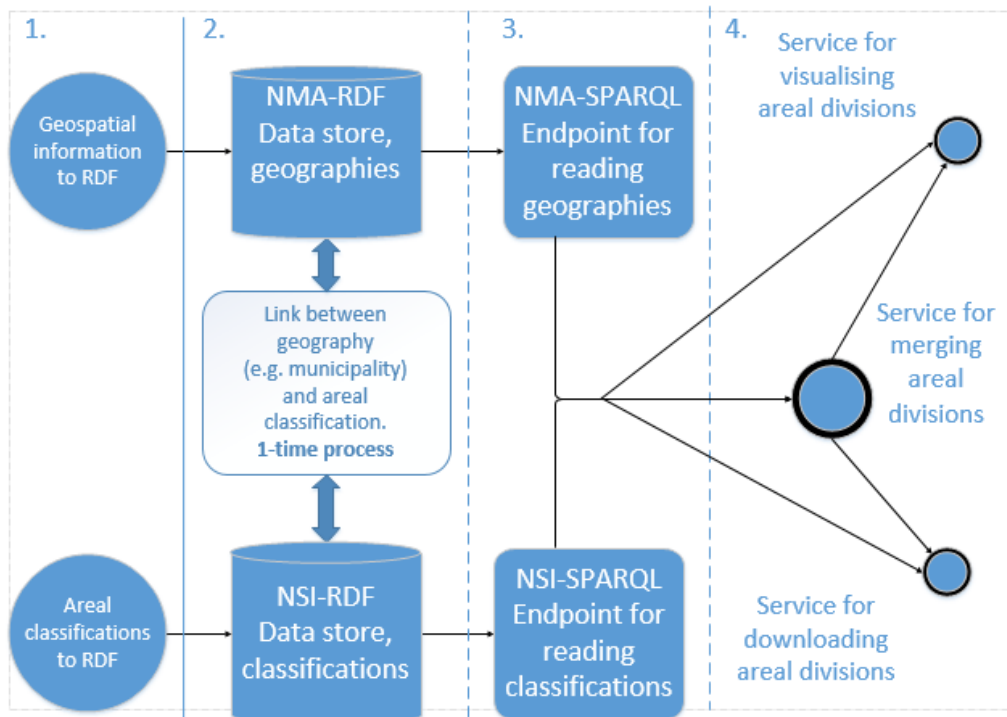


Figure 1: Suggested production model

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. However, the production model was partly simulated in a demo implementation and the results from that have been very promising. In the demo implementation both the classifications and the geographies were transformed into RDF and links were created between them utilising the unique identifiers of the data. For providing the data, only

one RDF store was used in the demo instead of two separate SPARQL endpoints. Also, in the demo implementation there was not a service for merging the geometries of the areas. However, the results were visualised on map in a free demo application available on the Internet.

Here is an example of a result when querying combination of geographic data, an areal class (municipality) and statistical data (Try it out: <http://yasgui.org/short/E7rL2s5sg>):

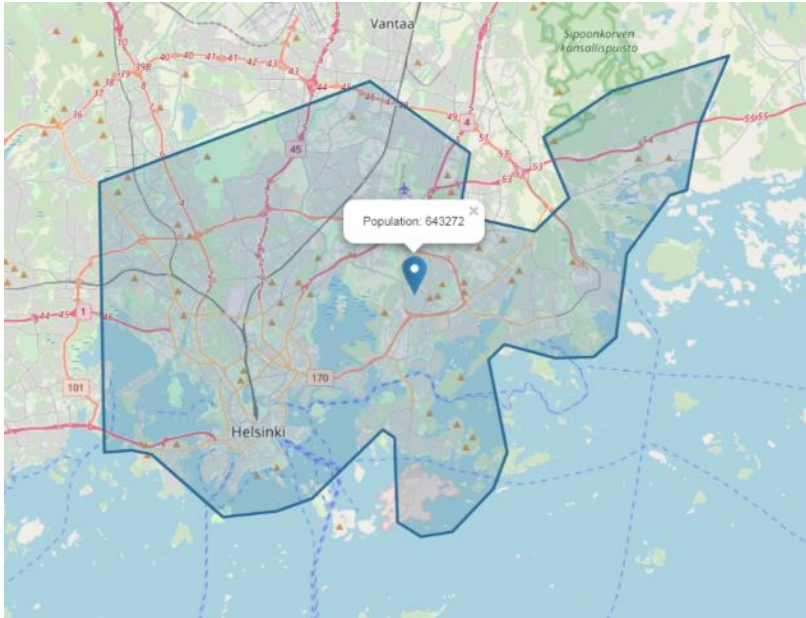


Figure 2: Example of querying combination of geographic data, areal class and statistical data

The same example using different parameters for geographic data to get more detailed geometry:

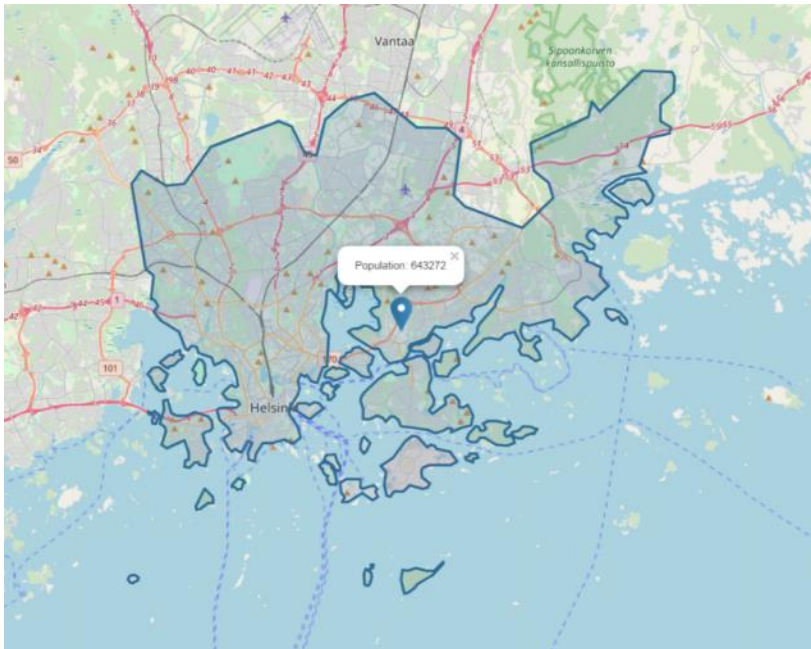


Figure 3: Example of querying data to get more detailed geometry

This is an example of a result when querying combination of an areal class (Region), included municipality classes according to classification key, corresponding geographic data and corresponding statistical data. (Try it out: <http://yasgui.org/short/sB98izjwM>):

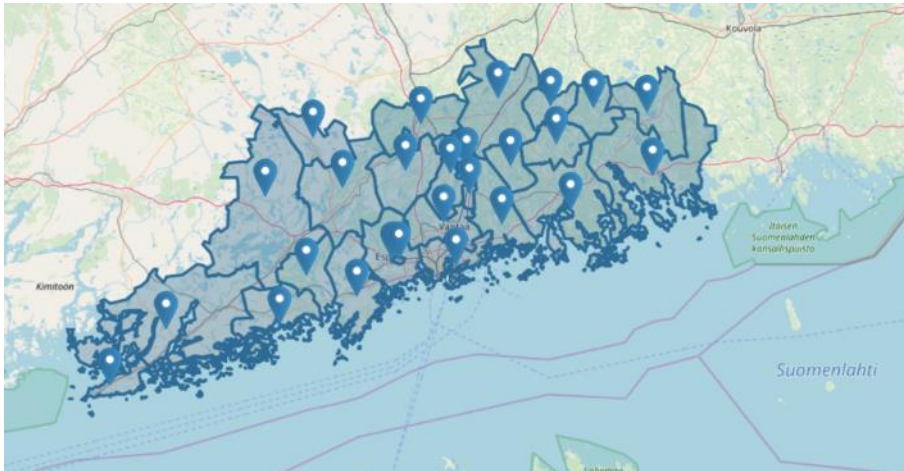


Figure 4: Example of querying municipalities according to classification key and corresponding geographic and statistical data

More information

There are not yet public documents available, but the link to the SPARQL demo service is here: <http://193.167.189.160/igalod/fuseki/>

The visualisation of the results was done by using YASGUI: <http://yasgui.org/>

Contact information

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

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

(C 4.5) Development of guidelines for publishing statistical data as linked open data (Poland)

Keywords: Principle 4, linked data, data integration, interoperability

This use case addresses multiple aspects of interoperability as requested by Principle 4, but especially those found in requirement 4.5, exploring the potential of Linked Open Data for increased interoperability. The use case demonstrates a solution but also addresses some of the challenges found with Linked Open Data. In addition, the use case briefly describes use of DCAT metadata as stated in requirement 5.4, facilitate data search and use through cataloguing and improved guidance.

Introduction

In January 2018 Statistics Poland concluded the “Development of guidelines for publishing statistical data as linked open data” project. The aim of the project was to perform a thorough inventory of data sources and investigate technologies, which could be used to publish georeferenced statistics as linked open data.

Data samples from statistical databases and geospatial datasets have been selected for transformation to linked open data RDF triples and a dataset catalogue has been set up and encoded in RDF. A pilot triple store has been established with a SPARQL endpoint – a query interface. Aside from the pilot’s results being machine readable, all data created in the pilot was also internally published as human readable webpages using linked open data frontend software.

Description of problem

The main goal of the project was to discover what types of data need to be transformed into linked open data and simulate a full statistical linked open data implementation on data samples. The data scope comprised three types of data:

Statistical data

A test scope of data from three major Statistics Poland’s databases (Local Data Bank, STRATEG, and Demography Database) has been selected for the purpose of the project. The data selected as a representative sample was the population by sex and age groups. Values for 348 variables were picked from the Local Data Bank, in the case of the Demography Database the number was 339, and in the case of the STRATEG database – 48. The selected data was used to design an age ontology for the aforementioned databases.

Samples of population by sex and age groups in 2016 for Poland and voivodships (the highest level of administrative division) were selected from the three above mentioned databases for conversion of statistical data to linked open data.

Spatial data

Statistical data chosen for publishing as linked open data contain a reference to territorial division identifiers. To make this data more usable, geometries of the territorial division units should also be published as linked open data. For purposes of this pilot, the geometries of Poland and voivodships have been selected for transformation.

Data source catalogue

The inventory of data sources carried out as one of the first stages of the project resulted in a structured catalogue of data sources described with metadata. Creating a dataset catalogue using linked open data was considered a valuable exercise for the pilot implementation.

Solution

The linked open data pilot implementation comprised two stages: designing ontologies for data and encoding data into RDF graphs:

Designing ontologies

The most important stage of designing a linked open data implementation is designing ontologies for published data sources. While it is at times necessary to design a whole new ontology for a data source, it is a good practice to re-use existing ontologies and vocabularies which have already been published. Publishing new ontologies which are similar to the ones already published should be avoided.

A thorough research of basic vocabularies as well as existing linked statistical data implementations is essential for designing useful ontologies. The core vocabularies are mostly stable and well described but during the course of the project it was very hard to find an implementation of statistical linked open data, which could be considered a reference. The biggest problem with existing implementations is that most of them have been published few years ago and have not been updated since. Not establishing repeatable processes in the organization to regularly revise and update the datasets results in most of the published data seeming abandoned, losing their value over time. Some of the implementations are internally inconsistent, e.g. it is apparent that different software tools have been used to publish different datasets and the resulting RDF files are not fully compliant with each other. Linking to other data sources is a vital part of linked open data. Sadly not all resources are being maintained and older implementations tend to have links to resources which are not available anymore. Nevertheless every implementation is a valuable source of information on how to model statistical data using linked open data vocabularies.

In terms of statistical data the pilot focused on publishing demographic data on population by sex and age groups. Age groups used in publishing data on demography are usually country specific, that is why it has been decided to create a new ontology for the age classification. For the sex dimension an existing SDMX codelist was sufficient.

In terms of spatial data the pilot focused on creating an ontology for the Coding System for Territorial and Statistical Units (KTS), which comprises territorial units used for dissemination of statistical data in Poland. The Open Geospatial Consortium (OGC) GeoSPARQL² standard was used to model relationships between classes and to encode geometries.

For cataloguing data sources identified in this project and datasets of statistical and spatial data, the DCAT Application Profile³ for data portals in Europe has been used. All datasets have been described

² <http://www.opengeospatial.org/standards/geosparql>

³ <https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe>

with metadata, re-using existing vocabularies where possible, e.g. EuroVoc for thematic categories, EU Publication Office Continent and Country lists or Internet Media Type (MIME) vocabulary.

Data encoding into RDF

Python RDFlib package was chosen as a tool for encoding RDF metadata. Main features of the package:

- tools to create RDF triples and store them in graphs,
- common namespaces such as RDF, RDFS or SKOS are already defined, other namespaces can be defined and bound to desired prefixes,
- parsing for RDF-XML, N3, NTriples, N-Quads, Turtle, TriX, RDFa and Microdata formats is possible, which allows transforming existing RDF metadata files to different formats, including RDFlib triples which show the exact syntax of the triple (subject, predicate, object). Parsing was especially useful for transforming files found on the Internet to learn how to construct triples and how they transform between different formats (e.g. RDF-XML and Turtle),
- serializing for all above mentioned formats is possible, so output files can be written in several formats.

The single most important advantage of the RDFlib Python package is the possibility to construct triples in any desired way (without the need for arduous configuration of mappings). Other advantages include the possibility to easily create output files in different formats and to modify Python scripts while working with the trial-and-error method or to supplement them later with links to other data sources or new vocabularies. All resulting RDF metadata in the pilot project was created using Python scripts.

Result

The pilot project was performed only on data samples, as it's main goal was to acquire knowledge on linked open data technologies and vocabularies. No data was disseminated to the public, all products were stored and tested internally.

Triple store and SPARQL endpoint

Apache Jena Fuseki software was used as a SPARQL server. Fuseki functions as both: a triple store and a SPARQL endpoint. All RDF graphs created within this project have been serialized and exported in both: RDF-XML and Turtle (TTL) format. Fuseki supports upload of both file types. All data was loaded into the triple store using the RDF-XML files. This project resulted in creating 71 717 triples. All triples have been loaded as a single Fuseki dataset to allow cross-querying and cross-browsing data created initially in separate files.

Linked open data frontend

Linked open data loaded into a SPARQL server allows data querying. Query results, which are URIs are provided as links. To make these links resolvable, a linked open data frontend needs to be set up. For this purpose Pubby software was used. Pubby creates webpages for each local URI defined in the datasets uploaded to the Apache Jena Fuseki SPARQL server. Each URI created within the pilot project

could be viewed with all its associated properties as a webpage, which allows browsing through all published linked data entities.

Conclusions

The project yielded valuable conclusions regarding statistical linked open data implementations:

- **No reference implementation for statistical linked open data.** There is no implementation of statistical data as linked open data that can be considered a fully correct, reference implementation. There are several existing implementations but most are plagued with some of the following issues:
 - lack of integrity between RDF metadata sets published by one authority – probably due to different software or programming components used,
 - links to non-existing entities – some implementations link to ontologies that have been published some time ago but are not online anymore,
 - lack of maintenance – most implementations are being developed and published but not later maintained.
- **Lack of pan-European guidelines for statistical linked open data.** Currently there is no guidelines for providing statistical data as linked open data (e.g. which vocabularies or software components to use),
- **Software / programming components not being developed anymore.** Some of the tools tested within this project (e.g. Pubby) are not developed anymore, so their implementations might become unstable in time. Python RDFlib package seems sustainable at this point (triples are produced based on encoding subject-predicate-object statements which are then serialized in stable formats like RDF, TTL), but it is also not developed anymore.
- **Not much data to link to.** Linked open data makes most sense if it is connected with as much other data sources as possible. This project utilized several existing vocabularies and already published datasets but a reference statistical linked open data implementation would be a much more desired resource to link to.
- **Semantic harmonization of statistical classifications.** This is not only a pan-European issue. It may also exist on the country level, if several datasets have different meanings for the same classification elements (e.g. difference in interpretation of age groups: 0-5 can be “0 to 5” or “0 to less than five”). Harmonization is always a difficult and complex issue, hopefully some conclusions in this matter will emerge from pan-European activities in statistical linked open data.
- **Methodology for publishing spatial data as linked open data.** This topic has two aspects: technological and temporal. In terms of technology, GeoSPARQL seems to be the correct way to publish spatial data as linked open data. The temporal aspect is much more complicated. In this project it has been decided to support publishing separate statistical unit geometries for respective years, regardless of their changes in time (due to different quality of spatial data for different years). The URIs have been constructed based on meaningful identifiers (KTS unit codes). A more appropriate situation would probably be a thorough analysis and inventory of statistical unit boundary changes in time and providing separate geometry instances with non-

meaningful identifiers (UUIDs). That would mean for example a single geometry with a defined period of validity for a unit which boundary did not change over the years.

- **Most linked open data implementations are technically correct.** By using existing software or programming components it is nearly impossible to produce incorrect RDF metadata files, regardless of the chosen encoding. The downside is, that most linked open data producing components allow encoding almost anything into triples, so the implementations may not always make sense semantically.
- **Linked open data implementations based on Python scripts are easy to amend in the future.** A big advantage in building a linked open data implementation based on Python scripts is their flexibility, which allows easy changes and amendments in the future.

Contact information

Mirosław Migacz, Statistics Poland, m.migacz@stat.gov.pl or kancelariaogolnaGUS@stat.gov.pl

5 Principle 5: Accessible and usable geospatially enabled statistics



(C 5.1) PX-Web API adapter for the Oskari platform (Finland)

Keywords: Principle 5, service oriented data dissemination, dissemination platform, dynamic mapping, API

This use case mainly refers to Principle 5, but some elements are also from Principle 4. Within Principle 5, this use case represents requirement 5.2: Use service oriented data portals supporting dynamic integration of data.

Introduction

The Oskari platform did not support PX-Web API (*Application Programming Interface*) as a data source for thematic mapping. Statistics Finland uses PX-Web databases, so for Statistics Finland's needs the Oskari's thematic maps function was not useful, since data from Statistics Finland's databases could not be presented there. The solution is an adapter that is compatible with statistical data cubes read from PX-Web API. The statistical data can now be read from PX-Web API and the data can be presented in thematic maps, tables and charts in Oskari.

Description of problem

Thematic maps function in Oskari was able to read only one specific data source, Sotkanet REST API. Statistics Finland is providing data from PX-Web databases, so for Statistics Finland it was a problem that Statistics Finland's data could not be read from a database to Oskari.

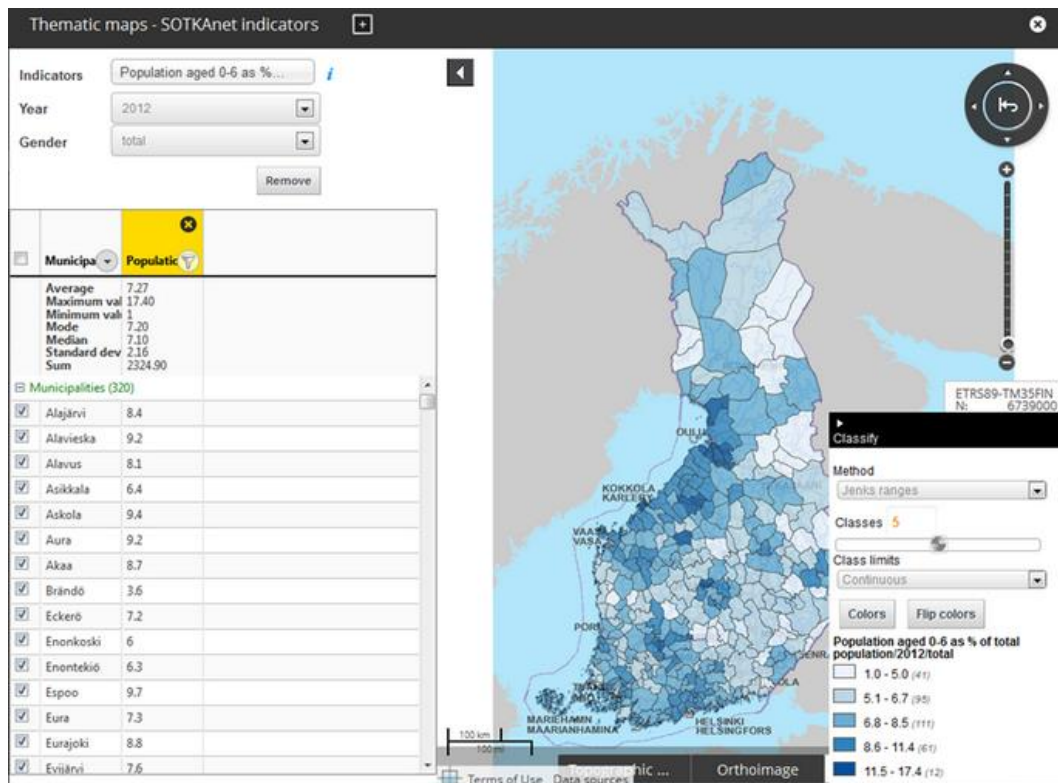


Figure 1: Situation in the beginning: only one data source for thematic maps

Solution

The solution to the problem was to refactor the thematic maps function in Oskari to be more general. The idea in the new implementation is to use adapters that make it possible to utilise any kind of statistical interface service.

For each type of interface service there will be a specific adaptor that reads and interprets the statistical data and combines it with the geospatial data. The geospatial data is coming from OGC interface service WFS (Web Feature Service).

Both datasets (statistical and geospatial) need to have some field like a region id that is used to join the two data. At the moment, Oskari includes statistical data adapters for SDMX REST, Sotkanet and PX-Web API interface services.

PX-Web API interface adapter

The PX-Web API endpoint is parsed recursively through any available “realms” to find all the available data “tables”. The table metadata is fetched to get any variables the table contains. The variable for region id needs to be configured for the API endpoint. For any time series functionality the time variable should also be configured. Any non-region variables are listed for the end-user in the user interface. Currently the adapter assumes that the region variable is shared between all the tables so the name of the variable is the same for all tables. Some additional metadata for the table should be generated on PX-Web for selecting the region set(s) that can be used with the table.

The adapter is responsible for turning the data from PX-Web API to the internal structure of Oskari. This means that the user interface does not need to care where the data came from. The transformation is done on the server side. The internal API requires the PX-Web API endpoint to be configured along with the region variable name. The code then generates a listing of tables available in that particular API endpoint (1). Once the user selects a table the variables (2) and optional description of the table is shown to the user. The region variable is not included in the end-user selections as it is used to match the statistical data to its geospatial counterpart (5). Any other variables are shown to the user including the option to select any geospatial region set the data is configured to be usable with (1). Once the user has defined the variable settings, the actual data is fetched from the live API (4). Everything else is cached once fetched.

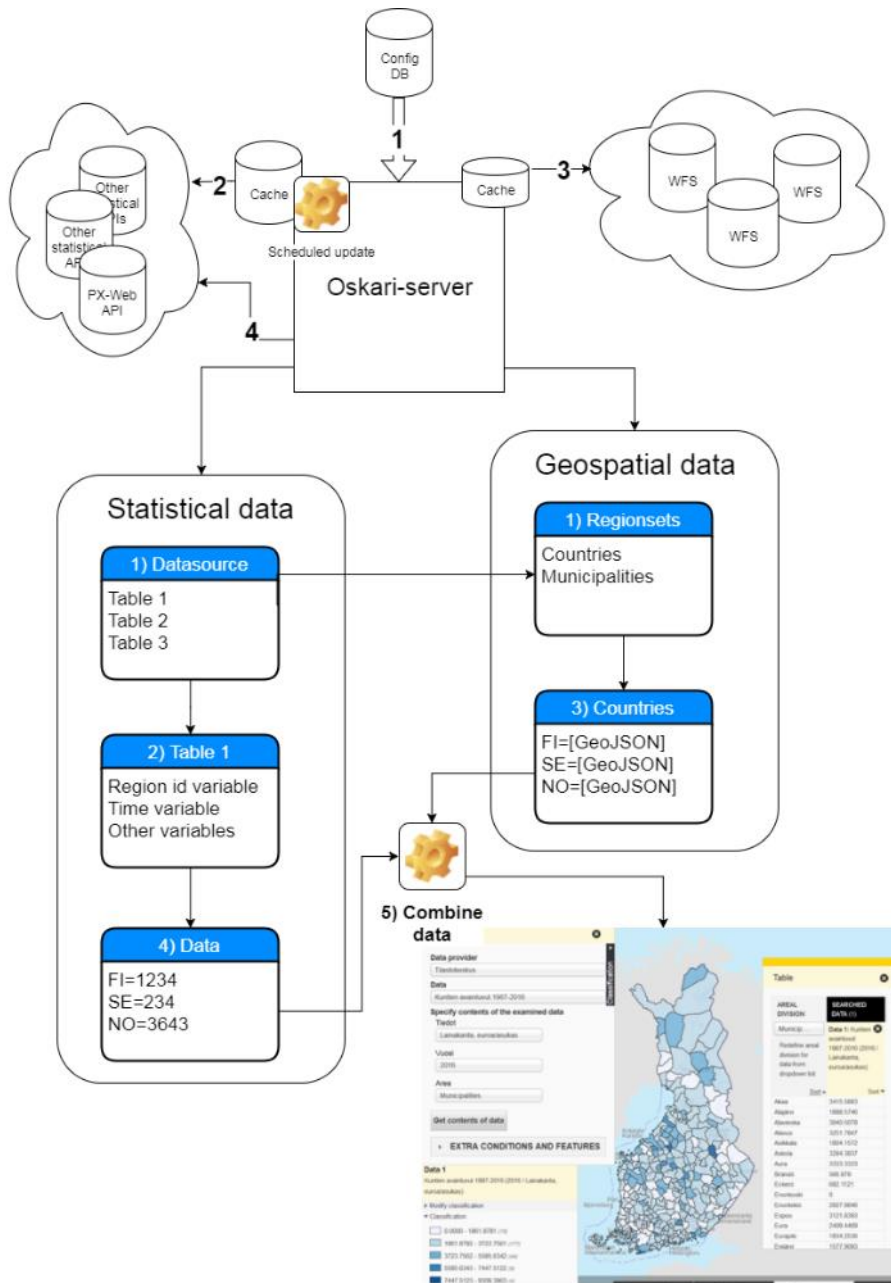


Figure 2: Solution of Oskari adapters for combining statistical data and geospatial data

Result

After the implementation of the PX-Web API adapter in Oskari, it is also possible to read data from PX-Web databases. Now Statistics Finland's statistics can be read from the PX-Web database and be presented in tables, charts and thematic maps in Oskari-based applications and dynamic maps.

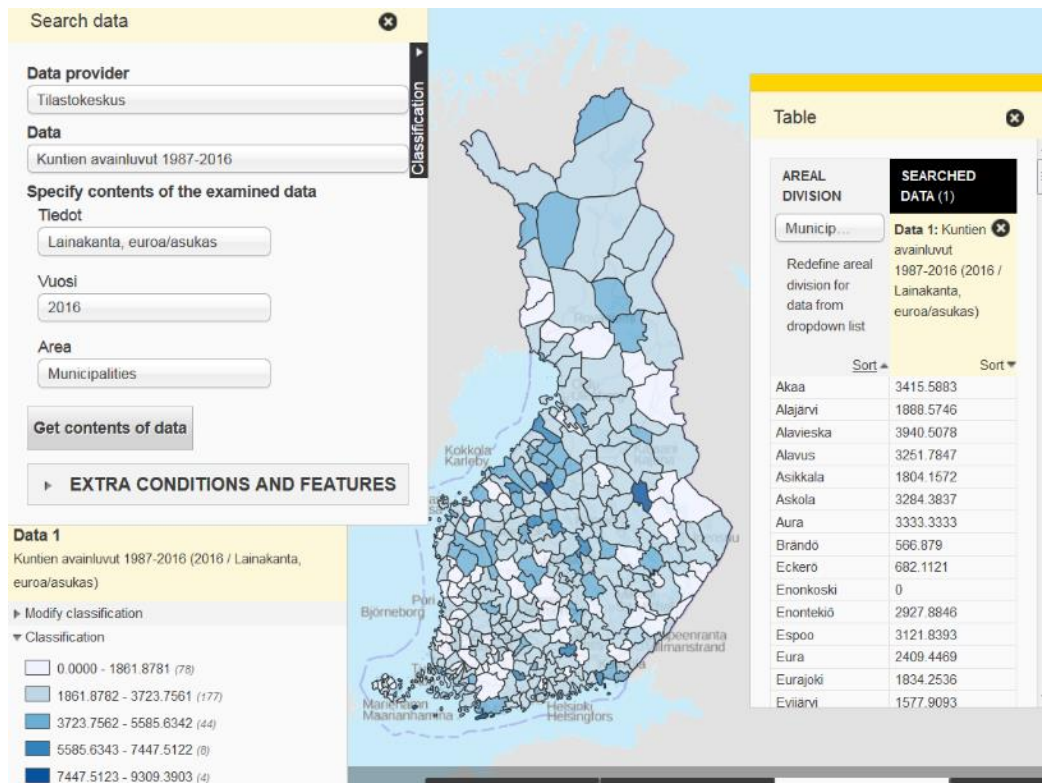


Figure 3: Situation after the new adapter: Statistics Finland (Tilastokeskus) as data source for thematic maps

More information

About Oskari: <http://www.oskari.org/>

Oskari demo platform: <https://demo-kartta.paikkatietoikkuna.fi/?lang=en>

The source code related for the PX-Web API interface service adapter in GitHub:

<https://github.com/oskariorg/oskari-server/tree/master/service-statistics-pxweb/src/main/java/fi/nls/oskari/control/statistics/plugins/pxweb>

Contact information

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

Contributors

Timo Aarnio, National Land Survey of Finland

Sami Mäkinen, National Land Survey of Finland

Tuuli Pihlajamaa, Statistics Finland

