**Integrating areal classifications and geospatial data**

*– case IGALOD*

Jennika Leino, Statistics Finland, jennika.leino@stat.fi

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

Essi Kaukonen, Statistics Finland, essi.kaukonen@stat.fi

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

Daniel Davis, Statistics Finland, daniel.davis@stat.fi

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

Kai Koistinen, National Land Survey of Finland, kai.koistinen@nls.fi

**Abstract**

*The importance of geospatial data in statistical production has increased recently. Geospatial information can have, in fact, different roles in different parts of the statistical production process as well as in different statistical programs. It can be source data for production, but it can be the outcome and final product of the process too. Quite often geospatial information is utilised as auxiliary data, while at other times it may, for example, enrich the statistical process with geospatiality.*

*A challenge in this area is how to actually integrate geospatial data in the statistical production process. One way to approach this issue is to link areal classifications and corresponding geographies using linked data methods.*

*Currently, the data is produced in two different organisations, geospatial data in the National Land Survey of Finland and classifications in Statistics Finland. The data structures are not harmonised and processes are separated and partially overlapping. In our solution, both organisations produce their data using common ontologies, and links between geographical data and classifications are described in the ontologies. Unique identifiers are used to identify the objects mapped together.*

*In our presentation, we share our experiences with linked data related technologies as well as the co-operation between Statistics Finland and National Land Survey of Finland on creating a common production process. Our technical solutions and decisions have been guided by common geospatial and statistical standards, so other national statistical organisations should be able to utilise this approach.*

**Keywords:** classifications, co-operation, geospatial data, linked data

**1. Introduction**

Geospatial information has been gaining more acknowledgement as an important part of the statistical production process over the recent years. However, the geospatial part of the work has usually been separated from the statistical production process. On the other hand, there has not been a common way to integrate the geospatial information in the production process either. In Statistics Finland there are in use international statistical process and information models (GSBPM and GSIM) and studies have been made, could they be utilised in the geospatial related statistical process.

In this conference paper we introduce a one way to integrate the geospatial information in the statistical production process. Our solution is to merge the areal classifications and corresponding geographical data together. This work is performed in co-operation of Statistics Finland, which provides the classification data, and the National Land Survey of Finland, which produces the geospatial data. We decided to use the linked data techniques to carry out this project and publish the data as open data.

In the section 2 we introduce the IGALOD project and the objectives of the project. Section 3 describes the solutions we have made to solve the issues of the project. Finally, in section 4 we draw together lessons we have learnt during this project and conclude this paper.

**2. Description of the project**

Two years ago, we had in Statistics Finland a chance to apply for a Eurostat grant on merging statistics and geospatial information in member states. We started the project, Integration of Geographies and Areal Classifications as Linked Open Data (IGALOD), in January 2018 in co-operation with National Land Survey of Finland. IGALOD is a two-year project and will be concluded by the end of 2019.

Statistics Finland (hereafter StatFI) and The National land Survey of Finland (hereafter NLS-FI) have been working together in other projects before, and this project continues the good co-operation. The management and main responsibilities of this project are carried by StatFI but all the tasks were planned and scheduled together.

This project was a good opportunity to create a way to enable the use of geospatial data in statistics production. In this project we could experiment techniques we had not used a lot before as well as how to unite the data production of two organizations.

*2.1. Objectives of the project*

The main objectives of the project are to create the links between areal classifications and geographical objects, publish the data as linked open data (LOD) and use the links in a new interactive map application. In linked data the persistent identifiers for the data objects are needed, another objective therefor is to implement an identifier system to manage URIs (Uniform Resource Identifiers) for areal classifications and geographical data.

The linking itself needs to be done within the data model, so the goal is to create ontologies for integrating the areal classifications and geographies. Relevant standards, like GSIM (Generic Statistical Information Model) and INSPIRE directive, need to be considered.

The third objective is to implement an RDF-based (Resource Description Framework) solution using the defined ontologies and identifiers. This solution will integrate areal classifications and geographies in practice and serve as a linked data repository for any possible applications.

As a part of the project we aim to pilot the outputs by visualising the links on an interactive map application.

**3. Solutions of the project**

This project includes multiple different kind of tasks that all together should lead to a single destination. The target is the ability to use data of two different organizations together to gain new information. We arranged the tasks of the project in five work packages (WP), of which the WP1 included the administrational tasks of the project. Other four work packages focused each on different task areas (Identifiers, ontologies, RDF-based solution and piloting). The work packages rely of each other and in practise they needed to be carried out partially at the same time. In the beginning of the project we decided to test our preliminary ideas on a Proof of Concept (PoC) as a cross-section of different tasks. In practise we transformed some areal classifications and geographical data into RDF and tested that data in an open-source map application. The results were encouraging, so we could continue to the chosen direction.

In this project it was important to take the relevant standards into account. The Generic Statistical Information Model (GSIM) (UNECE 2019 a) describes the information objects used in statistical production as well as their relationships to each other. The classification system in Statistics Finland is based on the separate GSIM Statistical Classifications Model (UNECE 2016), which is ingested into the model in the new version of GSIM. The Generic Statistical Business Process Model (GSBPM) (UNECE 2019 b) is used to describe the statistical production process. The INSPIRE directive obligates both StatFI and NLS-FI. In addition to these standards, the linked data paradigm is based on certain standards that need to be acknowledged as well.

*3.1. Geospatial data in statistical production process*

The geospatial data is an important part of the statistical data system in Statistics Finland. However, the geospatial data has traditionally been separated from the statistical production process and the possibilities geospatial data might bring, have not been fully appreciated. The reason for this was that necessary competence of geospatial technologies and administration has not been available. This has fortunately changed in the resent years and several actions in StatFI have been carried out to improve the case. Lots of work has been done as part of, for example, the GEOSTAT projects and now IGALOD project is putting its effort on it as well.

One of the objectives of this project is to identify the GSBPM (Generic Statistical Business Process Model) phases where the solutions of this project can be utilised. Geospatial data can be used throughout the production process and it has different roles in different phases. It can be source data and the starting point for production, but it can as well be the outcome and final product of the process. Or it can be used just to enrich statistical data with location information or with other geospatial elements, when the role is more supplementary. Basically, the geospatial data can be used in every phase of GSBPM (Kaukonen et al. 2017).

At present, the geospatial-related statistical production is based on point-base information, for example on the coordinates of buildings. To increase the use of geospatial data in the statistical production, a link between statistical and geospatial data is needed. In IGALOD project this link is created through the areal classifications. Like the geospatial data, the classifications can be used in different phases on the statistical production process too and they come naturally along with the statistical data. The contact point of the areal classifications of StatFI and the geographical data of NLS-FI are the municipalities. The NLS-FI produces the geometries for the municipalities, so combining them with the municipality classification enriches the statistical information.

*3.2. The Identifier Service*

One of the main issues related to linked data techniques, is how the data is identified. The Tim Berner-Lee’s (2006) four Linked Data principles note that:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs. so that they can discover more things

A URI (Uniform Resource Identifier) is a character string that is used to unambiguously identify resources on the web (Sauermann et al. 2007). It can contain information about the location and the name of the resource. Uniform Resource Locator (URL) and Uniform Resource Name (URN) are subsets of the URI (Fig. 1).

**Figure 1. URI, URN and URL (Sauermann et al. 2007).**



In Statistics Finland we have not had a common Identifier Service before, so the first task was to solve how to produce and manage identifiers for classifications. We started by creating a governance model for an Identifier Service, where the generic idea is to use GSIM-based information model TIMO to recognize all identifiable artefacts needed in statistical production process and in dissemination. These objects can then be added into the identifier system. The TIMO model is still in very early stages and will be developed further taking into account the needs of statistical production.

As the linked data principles suggest, the objects should be identified with URIs. The system is supposed to serve other production systems as well, so other identifier options are possible to insert later. For example, URN identifiers may come relevant for some other objects like publications and DOI for electronic documents, images, videos and so forth.

The system built is at this moment in its pilot phase. IGALOD project will add the created identifiers on the classification data and test them during this project. Any necessary changes needed based on this piloting will be implemented in the Identifier System during the year 2019.

*3.3. Ontologies*

The term ontology comes originally from philosophy and studies the concepts of *being*. In computer science it is used, in summary, to describe concepts, their definitions and relationships (Uschold and Gruninger 1996). The aim is to create *shared understanding* of some domain of interest to be able to communicate and interoperate. In linked data the ontologies are used to describes the resources and their relationships to other resources. RDF (Resource Description framework) is a way to describe the ontologies. RDF is a graph-based data model and W3 Consortium standard (Bizer et al. 2009).

In the beginning of the project it was planned to create an own ontology for the areal classifications. Linked data and RDF were quite new in Statistics Finland and studying more, XKOS ontology was found. XKOS (eXtended Knowledge Organization System) is an extension for the SKOS (Simple Knowledge Organization System) and is designed to describe classifications (DDI Alliance 2019). It is based on the Neuchâtel model that describes statistical classifications and variables. The Neuchâtel model underlies also on the GSIM Concept section on which the Classification system of StatFI is based on. Because of that point, as well as the fact that linked data ideology suggests using well-known standards whenever it is suitable to make it easier to link data together, we chose to use the XKOS ontology to describe the areal classifications. The municipality classifications as well as the NUTS (Nomenclature of Territorial Units for Statistics) classifications were chosen to be transformed into RDF format.

With the geographical data the case is different. There is no ready-made ontology solution for the municipality data. However, the INSPIRE data model Administrative Units has been translated into RDF ontology. INSPIRE is a European Union directive that became operational at 2007. It aims to harmonize the European Union spatial infrastructure and create policies and activities to enable co-operation and interoperability (European Union 2007). The INSPIRE Administrative Units ontology is very straightforward translation from the original model. The Administrative Units ontology doesn't contain for example support for geometries in different scales that is needed in the IGALOD case. Therefor a custom ontology was created for the geographical data.

The linking between the XKOS classification data and the custom geographical data is solved by simply adding the corresponding Administrative Units (AU) resource on the classification data and matching the municipality codes together. As said, the municipalities are the link between the areal classifications and geographical data. The municipalities are also mapped to all the other areal levels, so it is possible to generate the geometries for the upper level areas as well. The NUTS codes were also added on the data to point out the connection between national and international classifications. The Finnish classification Regions corresponds the NUTS level 3, Major regions corresponds the NUTS level 2 and classification Mainland Finland and Åland corresponds the NUTS level 1.

*3.4. Piloting the solution with a map application*

To test the solution in action, it was decided to get a map application to visualize the data interactively. A key feature was that the application should be able to utilise RDF data. The application we used in the PoC when we tested our ideas, was able to handle RDF data but could not combine other data formats which are also needed. The open statistical data of StatFI on the other hand is available in JSON format on open API. To get more use of the map application, we decided to look for other option.

We found a partner to provide a map application for us, and at this moment the development is in progress. The purpose is that the application reads the classification data from StatFI’s RDF API and geographical information from NLS-FI’s RDF API and federates the data to create the areas on the map. Then it combines some statistical data from StatFI’s PX-Web API and visualises that on the areas on the map. The map application was named as ALLUsion. The syllables AL and LU come from Finnish language and are acronyms for “alue” = area and “luokitus” = classification.

In the application the user can choose areas, for example multiple regions or just a single municipality, then the year for statistics and some statistical information for that area to be shown. In the Figure 2. the user has chosen the municipalities of South Savo. The population by a square meter is visualised on the map. The data can also be represented on a table, and in this case the table is visible as well. User can also choose whether to use the classifications and geographical data from the year 2018 or 2019, the scale of the data and include or not include see areas. In the browser session the user can name, save, remove and return own custom-made areas.

**Figure 2. The ALLUsion web application**

**4. Lessons learnt and conclusions**

In the IGALOD project our aim has been to create a new way to utilise geographical data in statistical production process and deepen the co-operation between Statistics Finland and the National Land Survey of Finland. We also wanted to test the linked data techniques which we had not used that much before. This project has been an interesting, cross-organizational action and we have had great possibilities to learn lot of new things.

*4.1 The identifiers*

We started to study about the identifiers by inviting an expert from the National Library of Finland to give us a lecture. We found out that there are different kinds of systems to identify objects and it depends on the use, as well as point of view, which to choose. In the IGALOD project it was decided to use linked data techniques and for that the choice was simple. As pointed out before, the linked data principles express that the URIs are the recommended identifiers. Looking from the identifier point of view, it is not that simple after all. When the identifiers are meant to be permanent and unique, the URI is not so reliable. The syntax of an URI can be exact the same than the syntax of URL, the Uniform Resource Locator (Fig. 1), so the user can never actually be certain whether it is just a web address or an identifier. In theory the URIs should as well be permanent but web addresses tend to break.

For this project we decided to form URI identifiers for the areal classifications and geographical data, as the linked data principles suggest. But because the Identifier Service is supposed to serve other production systems as well, the service is designed the way that other identifiers can be added and managed on it too. The studies around other ways to identify object are still going on in Statistics Finland in other projects.

*4.2. Linked data techniques*

One of the main points of this project was to use linked data techniques. In Statistics Finland we had only little knowledge of them, so we learnt a lot during this project. One of the objectives were to create an own ontology for areal classifications. We started by studying the existing ontologies and comparing them with our GSIM-based classification data. It was very important for us that the classification ontology would follow GSIM. We soon realised that there was not an ontology made of GSIM but same and same-kind of terms used in other ontologies.

Our first thought then was to adopt all the terms with exact match from other ontologies, and for the rest take the terms from GSIM. There is some well-known and widely used ontologies, for example DCTERMS (Dublin Core) or SKOS (Simple Knowledge Organization System), that describe data object we needed. By that time, we also discovered the XKOS (eXtended Knowledge Organization System) ontology. XKOS is an extension for SKOS and it is actually created to describe classifications. It includes the correspondence tables and is also able to describe the hierarchy of classifications too.

However, we found this ontology first a bit challenging because of some terms it uses from GSIM point of view. XKOS uses, for example, a SKOS term “concept” to describe the classification item. We thought it should be separated from the actual GSIM term “concept” which has a different semantic meaning in this context. From broader perspective, this is an issue which should be further elaborated while designing metadata system enabling discoverability and metadata driveness inside an NSO.

However, in this project after a few discussions with the XKOS creators and inside our project group, we came up with the conclusion that we should use the XKOS anyway. This is a logical solution especially for linked open data. There are some factors which favour using XKOS here. Firstly, we realised that the terms in XKOS are used to describe the ontology itself and should not confuse with the semantic meaning. In other words, for example the term “concept” in XKOS is a way to describe the objects that are included in the ontology (E.g. “In this ontology there is a “classification item” that is a type of concept”). In GSIM the “concept” is a part of the data model (E.g. “There are concepts, classification items and variables in this data model”). Secondly, in the linked data philosophy the purpose is to create shared understanding and using existing, well-known vocabularies it is easier for others to understand what our data is.

There has been prominent work going on for publishing geographical data as linked data like OGC's (Open Geospatial Consortium) and W3C's (World Wide Web Consortium) joint Spatial Data on the Web Interest Group (<http://w3c.github.io/sdw/>) and GeoSPARQL (OGC 2012). However, there is no ready-made ontology solution for modeling the municipality objects as such.

A custom ontology was created for the geographical data. INSPIRE AU terms were reused in the ontology and GeoSPARQL vocabulary was used to present the multi-scale geometry types. For presenting spatial object metadata RDFS, Dublin Core, SKOS, VOID and Schema.org libraries are used. A couple of custom terms had to be created also e.g. to provide the geometry with sea areas and without. The idea to provide multiple different geometric representation for a single spatial object resource proves to be a practical solution when using the data e.g. for visualisation purposes.

*4.3 Co-operation between StatFI and NLS-FI*

Statistics Finland and National Land Survey of Finland have years of experience of co-operation. The organizations have been working together for example, in the GEOSTAT projects before and at the moment the next project is under planning.

We have also realized that we operate in the same information ecosystems and that the data we provide need to correspond with each other to serve our customers. The IGALOD project suggests that the geographies offered by NLS-FI would be published yearly with unique identifiers corresponding to the areal classifications which have for long been disseminated on a yearly basis. This way, in the future, the areal classifications and geographies from previous years can be mapped to each other to support the use of historical data.

One practical solution invented in the project was to map the areal classifications and geographies in the database only in the municipality level. The other geographies can then be created using ontologies and correspondence tables in between different areal classifications. Responsibilities can also be a lot clearer than they were before when this solution is utilized. NLS-FI needs to govern only the geographies of municipalities and there is no overlapping work done for the other layers.

In this project the project group was small but efficient. From StatFI there were five members and from NLS-FI two members. In the steering group there were three members from StatFI and one member from NLS-FI. Our co-operation has been very fruitful, and we have learnt from each other. Because of a small group, the atmosphere has been free and easy. The differences between the organizations has mostly been only in the terms we use in our work conversations. This is natural, because of the different sectors we are working on. At first, we had to learn the terms each organization uses, but after that it has been easy to communicate.

The management of this project has been Statistics Finland’s responsibility, but we have planned the work together. We have organized several workshops to work on some lager issues or for example, Skype-meetings to discuss smaller matters of the project. We arranged also a study trip to INSEE, the statistical organization of France, on which participated members from both StatFI and NLS-FI.

*4.4 Conclusions*

The main objectives of the IGALOD project were to link areal classifications and geospatial data together using linked data techniques, publish them as open data and pilot in a map application. It was important to make sure the relevant standards (GSBPM, GSIM, INSPIRE, etc.) were taken into account on the project. The project was carried out in co-operation of Statistics Finland and the National Land Survey of Finland.

We chose to transform the classification data into RDF format using the XKOS ontology. For the geospatial data a custom-made ontology was created. Also, unique identifiers for the data objects were given. Both organizations set up their own dissemination environments and released their data as open data. A map application was purchased to pilot the use of RDF data and to visualize them on a map.

In this project we have illustrated a one way to unite statistical classifications and geospatial data and production processes of two different organizations. As our results show so far, it is possible to produce and manage the data in different organizations and use them on-line. To take this solution into production needs more discussions in both organizations. The production and maintenance of the linked open data and the dissemination environments will need proper management and maintenance. This requires resources.

So far, this project has been very interesting, and the tasks have been completed with success. One of the most important outputs is also the fact that we have got a lot of new knowledge and competence.

**5. References**

Berners-Lee, T. (2006), Linked Data. Available at: <https://www.w3.org/DesignIssues/LinkedData.html> (Accessed: 29 April 2019).

Bizer, C., Heath, T. and Berners-Lee, T. (2009), Linked Data – The Story So Far. Available at: : <http://tomheath.com/papers/bizerheath-berners-lee-ijswis-linked-data.pdf> (Accessed: 24 June 2019).

DDI Alliance (2019) XKOS An SKOS extension for representing statistical classifications. Available at: <http://www.ddialliance.org/Specification/XKOS/1.2/OWL/xkos.html> (Accessed: 20 June 2019)

European Union (2007), INSPIRE Knowledge Base. Available at: <https://inspire.ec.europa.eu/about-inspire/563> (Accessed: 20 June 2019)

Kaukonen, E., Tammisto, R., Pihlajamaa, T. and Leino, J. (2017). Towards connecting geospatial information and statistical standards in statistical production: two cases from Statistics Finland. Available at: <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2017/mtg3/Towards_connecting_geospatial_information_and_statistical_standards_Statistics_Finland.pdf> (Accessed: 25 June 2019)

OGC, Open Geospatial Consortium (2012), GeoSPARQL - A Geographic Query Language for RDF Data. Available at: <http://www.opengeospatial.org/standards/geosparql>

Sauermann, L., Cyganiak, R. and Völkel, M. (2007), Cool URIs for the Semantic Web. Available at: <http://www.dfki.uni-kl.de/dfkidok/publications/TM/07/01/tm-07-01.pdf> (Accessed: 14 June 2019).

UNECE (2019) a, GSBPM v5.1. Available at: [https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1](https://statswiki.unece.org/display/GSBPM/GSBPM%2Bv5.1) (Accessed 25 June 2019)

UNECE (2019) b, GSIM. Available at: <https://statswiki.unece.org/display/gsim/> (Accessed: 25 June 2019)

Uschold, M. and Gruninger, M. (1996), Ontologies: Principles, Methods and Applications. Available at: <http://www.aiai.ed.ac.uk/publications/documents/1996/96-ker-intro-ontologies.pdf> (Accessed: 14 June 2019)