Towards connecting geospatial information and statistical standards in statistical production: two cases from Statistics Finland

Essi Kaukonen, Rina Tammisto, Tuuli Pihlajamaa, Daniel Davis, and Jennika Leino

Statistics Finland, firstname.lastname@stat.fi

Abstract: The Nordic countries have a long tradition of including geospatial information in the statistical processes, and having access to geocoded administrative sources is an important foundation. In more recent years, closer cooperation with the mapping authorities increases the possibilities to create more efficient processes and new statistical products.

However, geospatial information is not a clearly recognised part of the statistical production process. The geospatial part of the work is usually separated from traditional statistical work. The GSBPM and the GSIM statistical models, that provide a generic framework for statistical production, are in use in the Nordic countries. Could these models help build the bridge between geospatial information and statistical production? How could this be done?

Statistics Finland has investigated how the GSBPM and the GSIM could be utilised in the geospatial related statistical process. This paper will highlight some results and give examples of future work from Statistics Finland. Case A deals with the GSBPM and how this model could be used in describing the geospatial-related statistical production process. Case B introduces our future plans to connect the GSIM Classifications and corresponding geographies for areal classifications.

1 Background: The role of geospatial information and UNECE statistical standards at Statistics Finland

Geospatial information is getting more and more acknowledged as an important part of the statistical production process. This has recently been recognised in a project tasked with modernising the production of social statistics. In the following picture (Fig 1.1) you can see the logical information contents related to the person where the “Location information”-component is, indeed, common for all statistical fields.

On the other hand, the statistical standards governed by UNECE also play a significant role at Statistics Finland today. The Generic Statistical Business Process Model (GSBPM, Fig1.2) is included in our 2016 to 2019 strategy as the reference model for our core business process. The GSBPM was officially adopted as the framework for statistical production in 2015 as such; no national changes have been made to this model. Respectively, the Generic Statistical Information Model (GSIM) has been used since 2015 in some national and international projects. The GSIM Statistical Classification Model was chosen as the base model for our new Classification System, which has been running for over a year now. The overall role of the GSIM in our emerging Information Architecture is currently under discussion.

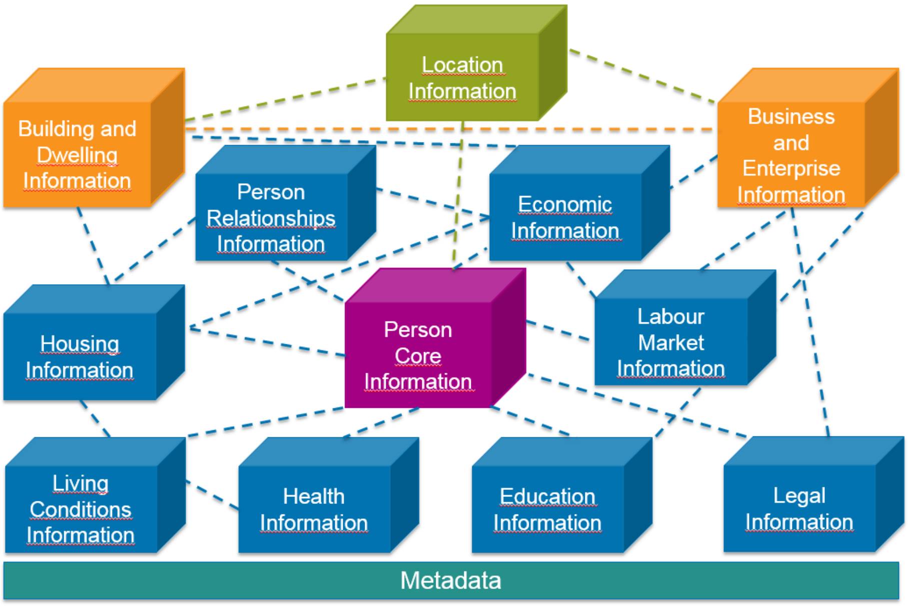

Fig 1.1 Logical information contents: social statistics.

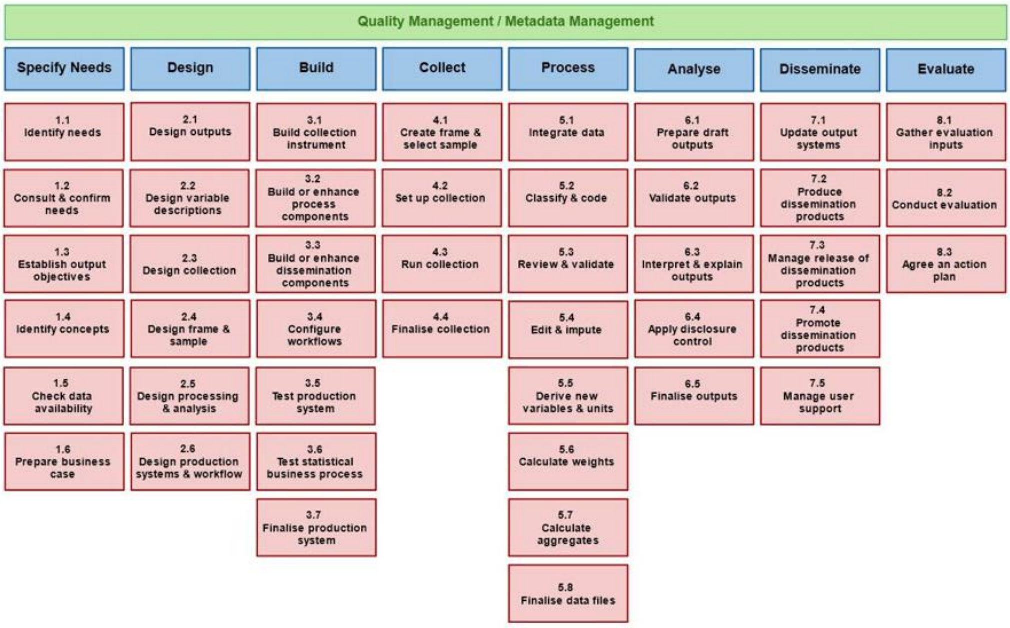

Fig 1.2 Levels 1 and 2 of the Generic Statistical Business Process Model.

2 Case A) Testing the GSBPM in the geospatial-related statistical production process

Geospatial information is getting more and more acknowledged as an important part of the statistical production process. The GSBPM is the model to describe this statistical process, but how could we include geospatial information in it? This has been one of the topics for a grant funded European project called GEOSTAT2.

The extension of the GSBPM model to recognise the use of geospatial information in the statistical production process would provide a much needed link to internationally agreed statistical processes and facilitate the communication between the statistical and geospatial communities. In addition, including geospatial information management in the GSBPM would extend the infrastructure for geospatial statistics to actual data production and, thus, bring it to the core business process of statistical offices. The GEOSTAT2 project concluded that UNECE should take this work further under the High Level Group on Modernisation of Official Statistics.

2.1 Background

One task of the GEOSTAT 2 Essnet project was to test and evaluate the Generic Statistical Business Process Model (GSBPM) when geospatial data are involved in the statistical production process. The questions were: 1) is the GSBPM a usable tool to understand and describe the geospatial dimension of the statistical production process? 2) if yes, does the GSBPM need further development or does its documentation need more accurate expressions when geospatial data are possibly included in the process? The project aimed to achieve common views of subjects that could be handed to UNECE for evaluation.

The starting point for testing was that the GSBPM has been implemented in many NSIs but is typically not applied to describe and structure processes implying the use of geospatial information. As there are no references to geospatial data in the GSBPM's current guidelines, there was no given baseline to test the model in a comparable way. Statistics Finland evaluated the usefulness from the point of view of current geospatial data production.

2.2 Geospatial data in the statistical production process

Geospatial information can be involved throughout the whole process or it may be only part of one particular phase of the production. So the role of geospatial information varies. Geospatial information can be source data and the starting point for production, but it can also be the outcome and finalised product of the process. When it is used to enrich statistical data with location information or with other geospatial elements, geospatial information has a role of auxiliary data. In any case, when location is an important part of the statistical production process it brings an additional dimension to the statistical production process and induces special demands and challenges to be solved.

A production process where geospatial data is involved may require special skills and technical environments at different stages of the process. Both statistical and geospatial experts are usually needed. It should be noted that if processes are modernised in a systematic way, automated or service-based production may reduce the need for different types of professionals. An example of the input of different specialists at different process phases are illustrated in the following picture. Certain phases may need more profound geospatial expertise than others. The geospatial expert is visualised in the figure 2.1 by a person with green background. In Process 1, geospatial expertise is dominant. In Processes 2 and 3, the geospatial issues are parts of one or two particular process phases. The spatial related work phases may appear only once but the impact of that phase may be crucial for the success of the process from the point of view of bringing spatiality to the production of statistics.

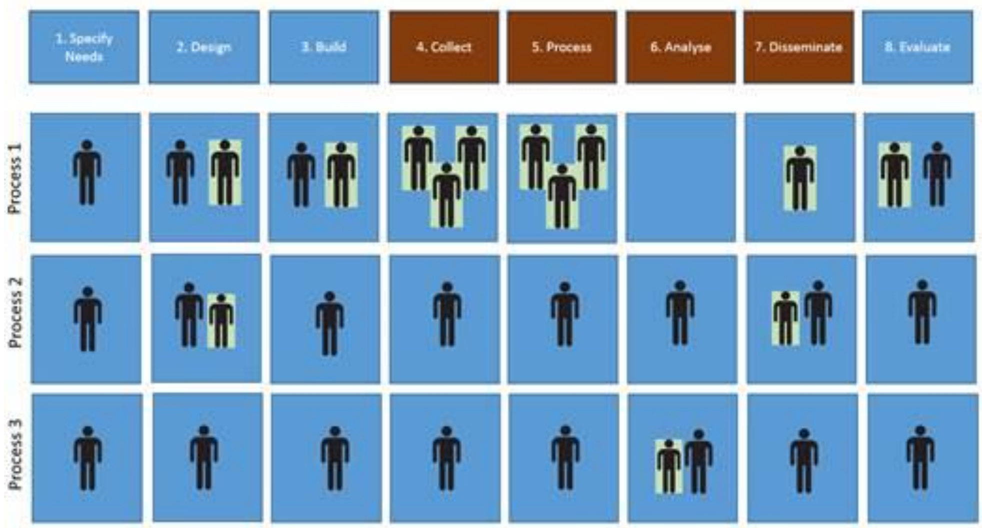

Fig 2.1 The need for geospatial data and expertise varies in the geospatial statistical production process.

When the experts of the geospatial and statistical sides are cooperating, a common language is needed. The GSBPM, as a common framework, offers the possibility to describe the production process in a way that mutual understanding is possible can be achieve. Best practices, instructions and tools can also be better shared. From an organisational point of view, the GSBPM offers a tool whereby separate processes are also able to communicate with each other. The geospatial dimension may be carried out in the same way or with a common tool in different processes. This is excellent not only in terms of efficiency, but also in terms of quality management, while a quite common situation is that stability management between processes is lacking.

2.3 Testing method and results of the tests

Geospatial-related production processes are not clearly connectable to the GSBPM phases, because there are no references to geospatial data in the GSBPM or in its guidance. Tests were made at Statistics Finland with an expert who is specialised in modernisation of statistical production processes. With his guidance, tests were performed by raising the abstraction level of the GSBPM model and the related documentation. The idea and the mission of a particular GSBPM phase was evaluated from the geospatial view point.

Two main points were made: 1) from the point of view of process phases, geospatial data may be seen as any other data. The necessary edits and other preparatory acts are equivalent to ordinary statistical source data modifications that are needed, thus geospatial-related work phases may need special preparations, arrangements or GIS technology, 2) work phases related to editing geospatial data can, almost without exception, be seen as a different process or as a sub-process under the main statistical production process. In Statistics Finland's tests, it was, therefore, decided to divide the process into two separate process flows (Fig 2.2). The first process flow describes the geospatial-related statistical process as a whole, the other describes the process relating directly to geospatial data. These two process flows cross at particular points, since the geospatially related phases are often necessary in a specific stage of the main statistical geospatial process.

The most surprising test result was that all geospatial-related production processes could fit into the same process flow (regardless of the content of data or geospatial character of data).

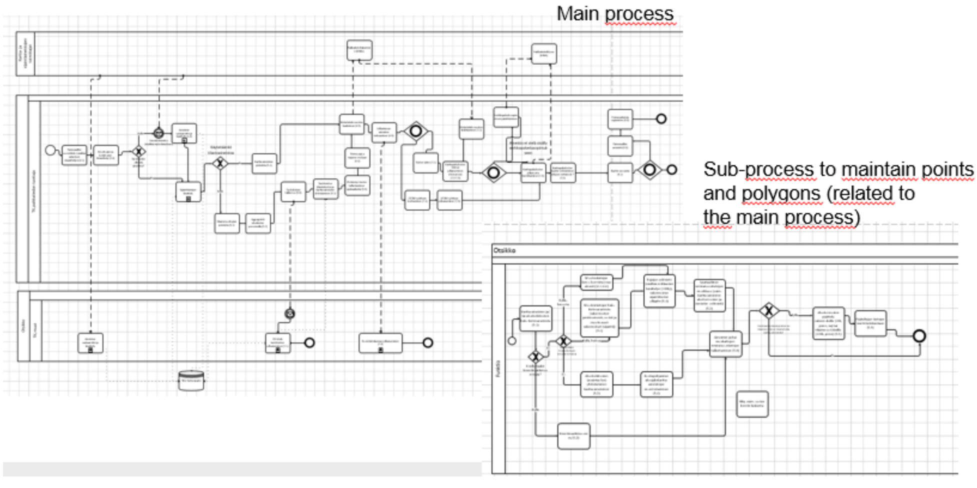

Fig 2.2 An example of a statistical production process where geospatial data are involved. The main statistical production process was separated from the elaborated process of

maintaining purely geospatial data. The main statistical production process is linked to geospatial-related work phases occasionally when needed.

The GSBPM phases also cover the purely geospatial process phases. This is possible if we interpret the logical meaning of the particular phase and apply its action to concern geospatial data. The integration of statistical data and geospatial data can easily fit to the GSBPM if integration of geospatial data and statistics is seen equivalent with the phase where two statistical datasets are joined (GSBPM 5.1 Integrate data). The same GSBPM phase may occur many times in a statistical production process, but also in case of geospatial data.

2.4 General remarks of the GSBPM model from a geospatial point of view

On phase level, geospatial use could be identified in all phases (Fig 2.3). We recognised 30 geospatial-related sub-processes in the 44 GSBPM sub-processes. In some cases, only a wider angle may help include geospatial. The concern thus is that if it is not specifically mentioned, its inclusion is not recognised.

Phases 1 to 3: The first three phases already demand certain specific understanding of the character of geospatial data. Evaluating, designing and building the statistical system require to understand demands and possibilities that the geospatial data brings. These may concern spatial dimension of data or special technology they require.

Phase 4: From Statistics Finland's point of view, geospatial data collection is usually register-based. If location information is not part of some register-based data collection it is very likely to be downloadable from the web service of some other authority.

Phase 5: The most concrete phase where geospatial and statistical data are integrated or geospatial data brings additional benefits is the process phase. This phase culminates the value of geospatial data to the statistical production process.

Phases 6 and 7: In the analyse phase, the location-based analyses take place. They may be needed for creating a new variable or a visualisation for the output dataset. In geospatial data dissemination, the geospatial standards offer a solid and uniform base for all statistical data providers.

Phase 8: Evaluation of geospatial-related process, data or services may need a special angle and technology to be carried out properly.

Phase	Remarks from StatFI tests	Other remarks
1. Specify Needs	The geospatial dimension of statistical data should be recognised and taken into account. When planning outputs also geospatial data, geospatial statistics or map visualisations should be noted. In addition, geospatial concepts should be taken into account next to statistical concepts. In the "Check data availability" sub-process also integratability with geospatial data should be evaluated.	ABS suggests 1.5. Check data availability should be broken into a new sub-process such as: 1.6. Check data suitability (incl. assessment of spatial suitability)
2. Design	Geospatial standards should also be noticed. When designing geospatial statistical data output, the scalability of disclosure control should be taken into account. When designing variable descriptions, the existing link between statistical classification and corresponding geographical data should be noted and described. Design collection should also include mention of techniques for managing geospatial data, e.g. mobile technology. When routines to integrate datasets are designed, integration of statistical data and geospatial data should be noted.	ABS also reminds that the platforms /systems available for dissemination must be considered. In particular, if visualisation of outputs in maps is desired, systems to enable this must exist or be developed. Also, designing of collection lacks mention of variables, which are generated directly from geospatial datasets.
3. Build	In building and testing the production solution, geospatial data and services that are used should have a general and stable basis, not separate solutions for different production processes. When geospatial data is the case, data gathering usually concerns use of an existing web service (WMS, WFS). Direct connections or data storage procedures may be used. The data collection system should be able to handle geospatial data.	ABS would include geospatial metadata such as the geometry of the unit record
4. Collect	Geospatial data may be collected from different data sources, providers and by using several different methods (web services or other services, e-mails, register based data from (other) statistical system. Collecting of geospatial data can be done manually or automatically.	
5. Process	This phase is the most concrete phase where the integration of statistical and geospatial data is present. It may take place by enriching the statistical data by the location information or it may mean to integrate the geospatial dataset to the aggregated dataset (integrate statistical data to corresponding boundaries). This phase also includes maintaining geospatial data with updated areal classifications or vice versa. This phase includes study of location information and identification of locational errors. Editing of geospatial data may consist of merging geospatial elements or creating new or changing the location information of spatial objects. During this phase, aggregation of data by areal classification, storing aggregated, integrated, edited or conducted data to data warehouses also takes place.	ABS mentions that geospatial data may be the only method for integration, for example using unit record address geometry to calculate a distance to a geometry for a service provider (access to services).
6. Analyse	This phase includes accessibility measures, travel and time distance calculations. Map visualisations are included in this phase. It should be kept in mind that disclosure control for geospatial data depends on the scale of geospatial data. Compiling metadata also for geospatial data is important.	ABS would also include spatial data products such as grids here. ABS and Statistics Poland remind that map visualisation needs a special spatial process. Statistics Poland would prefer an own sub-process for map visualisation actions here. ABS sees the protection against geographic differencing is important. ABS reminds that conformity to Geospatial metadata standards for geospatial datasets is important (ISO 19115)
7. Disseminate	It is important that geospatial data outputs conform to the geospatial data dissemination standards (ie. ISO, OGC, INSPIRE). The creation of geospatial web services (WFS, WMS) are included here. Instructions for use for users should also be included.	
8. Evaluate	Gathering information for evaluation of the geospatial process, data or created services, evaluation of this information and action plan for further development.	

Fig 2.3 Remarks on the suitability of the GSBPM for geospatial-related statistical production at Statistics Finland. Other remarks consists of comments received from the Australian Bureau of Statistics (ABS) and from Statistics Poland.

2.5 Conclusion Case A

One of the main benefits of using the GSBPM is that it helps to see where and how geospatial information might appear, or should appear, in various phases of the production process. The GSBPM can be seen as a tool to help NSIs manage the use of geospatial information in their statistical production process. Also, in geospatial related production, it makes geospatial elements and their role in the statistical production process visible. It makes requirements and possibilities visible. It offers a common language and a tool for a common understanding of the process. It may also increase quality and efficiency, especially between separate processes.

The GSBPM also appears to be suitable for describing the production of geospatial statistics. However, in order to recognise the GSBPM phases, plenty of interpretation is needed first because the GSBPM documentation does not include geospatial related instructions. The level of abstraction that is essential for understanding the process must be chosen. Raising the abstraction level of the GSBPM's documentation and adding references to geospatial issues helps locate process phases for each phase of the GSBPM. It is challenging to settle on a common way to use the GSBPM model, since the use of the model must be agreed on prior to common utilisation.

The extension of the GSBPM model to recognise the use of geospatial information in the statistical production process would provide a much needed link to internationally agreed statistical processes and facilitate the communication between the statistical and geospatial communities. In addition, including geospatial information management in the GSBPM would extend the infrastructure for geospatial statistics to the actual data production and bring it to the core business process of statistical offices. The GEOSTAT2 project concluded that UNECE should take this work further under the High Level Group on Modernisation of Official Statistics.

3 Case B) Linking the GSIM statistical areal classifications and corresponding geographies

Statistics Finland's new classification information system based on the GSIM Statistical Classification Model includes relevant classifications and code-lists used in statistical production. The areal classifications can be brought up, as any other classification, to serve all phases of the statistical production process (GSBPM) in all relevant systems.

On the other hand, geographies of the most common areal classifications have been produced at Statistics Finland and they are also published as open data using standardised service interfaces (WMS/WFS) on our webpages. This has been done as part of the INSPIRE implementation at Statistics Finland.

Currently, geospatial-related statistical production is mostly based on point-based information, for instance coordinates of buildings. To extend the use of geospatial data in the statistical production process, a link between the areal classifications and the corresponding geographies is needed.

In the following chapters, we describe both of these items to be connected. Future plans concerning merging the areal classifications following the GSIM and corresponding geographies are discussed on a concrete level as well.

3.1 Classifications in the new classification system at Statistics Finland

3.1.1 The new classification system for the statistical production process following the GSIM

Statistics Finland has renewed the classification information system originally built in the early 1990s. Classifications have been stored in the new system since spring 2016 and they can be utilized from it throughout the statistical production process. Currently, custom solutions for managing classifications also exist, especially in the middle part of production, GSBPM phases 5 and 6, but gradually, all statistical systems will be changed to use the central system.

The new system is based on Service Oriented Architecture (SOA) principles and thus can serve the statistical production process flexibly. SOA is a software design style in which basic principles are to provide independent, flexible and open services that can be used through interfaces. In the classification system, the classification data are stored in a relational database and used only through services; the Service Interface is a REST (Representational State Transfer) interface. The services provide requested information in XML or JSON formats.

The dissemination of classifications as open data is going to be accomplished in a development project starting in autumn 2017.

3.1.2 GSIM Statistical Classification model the Finnish way

The GSIM was chosen to be the model to be used in the new Classification System after careful analysis and testing. The model served the recognised needs of Statistics Finland.

The old classification system contained Statistical Classifications and Code Lists and we decided to include them both in the new system as well, for these are both needed in the statistical production process. During that time, the previous Neuchâtel Classification Model (Fig 3.1) was merged into the GSIM as the GSIM Statistical Classification Model.

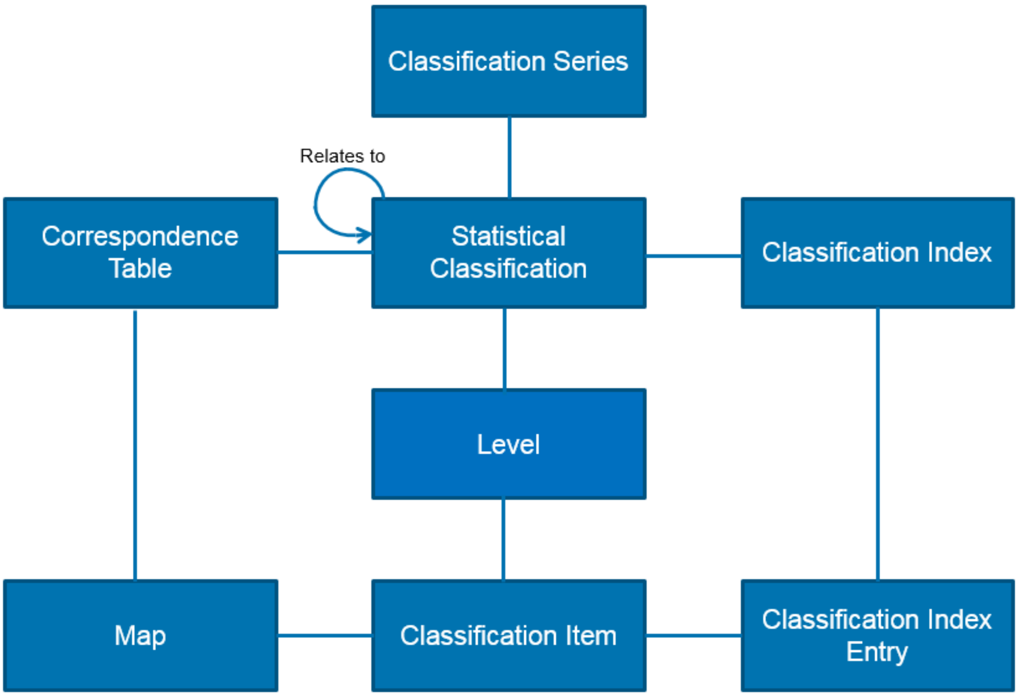

Fig 3.1 GSIM Statistical Classification Model, core elements.

In our new system we have implemented the GSIM Statistical Classification Model in a way that both Statistical Classifications and Code Lists are stored as Classifications, and the user can tag if the Classification is a "true Statistical Classification". The international and national Statistical Classification Standards can be tagged as well. All Classifications and also Code Lists belong to the Classification Series in our solution.

We did not implement the Node Set structure of the GSIM v.1.1 as such, mainly for three reasons:

• This structure was added into the GSIM at a time when we had already tested the separate GSIM Statistical Classification Model for both Statistical Classifications and Code Lists

• We have a decentralised governance model for classifications and, therefore, reusability of all elements like Levels and Classification Items was not reasonable

• The Node Set structure can be interpreted in practice in a way that the user has to make a clear distinction whether he is creating a Statistical Classification or a Code List. This might be a challenge as the variants of Statistical Classification Standards are, actually, quite often Code Lists, and the model does not support the idea of including Statistical Classifications and Code Lists into the same Classification Series.

3.1.3 Areal classifications in the new system

Areal classifications, like municipalities and counties, are stored in our new information system following the GSIM Statistical Classification Model like any other classification. This means, in practice, that a certain municipality is stored as a Classification Item. For instance, the capital city of Finland in our municipality classification is stored as a Classification Item, its Code being "091" and Name "Helsinki" (Fig 3.2).

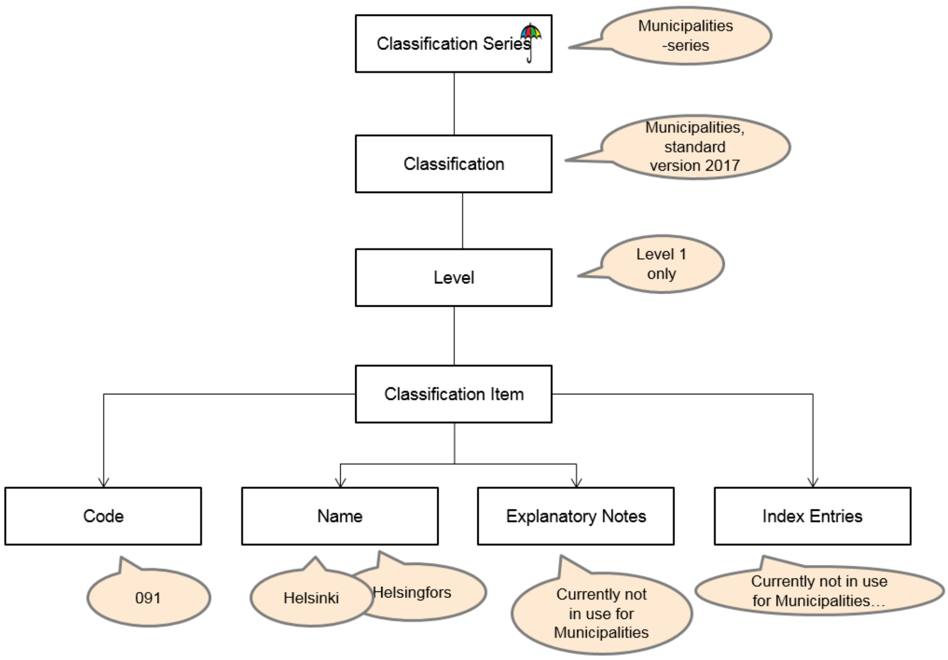

Fig 3.2 Areal classifications, an example.

3.2 Geographies and unique identifiers

3.2.1 Production of geographies and their publishing as open data

Statistics Finland is a data provider of INSPIRE data products for several themes. One of the data products is the theme Statistical Units, which includes municipality-based statistical units and grid net for statistics 1 km x 1 km.

Municipality-based statistical units are produced in Statistics Finland based on the geographical source data that is provided by the National Land Survey of Finland (NLS) and the areal classifications of Statistics Finland. Statistics Finland publishes geographies of the most common areal classifications as open data. INSPIRE recommended OGC (Open Geospatial Consortium) standardised web services (WMS/WFS) are used as the service base. Published geographies are not utilised in dissemination of areal classifications or statistical datasets.

3.2.2 Unique identifiers of spatial objects; INSPIRE URI recommendations

INSPIRE guidelines include recommendations for the unique identifiers of spatial objects. For example "Guidelines for the encoding of spatial data" and "INSPIRE Generic Conceptual Model" give instructions about the type and format of URIs. Among the most important recommendations are:

- It is recommended that identifiers of resources should be URIs in the "http" scheme.
- It is strongly recommended that resource can be accessed via the HTTP protocol, using its http URI.
- URIs of spatial objects should be persistent http URIs and they should include the namespace and the local identifier part of the INSPIRE identifier, if available.

3.2.3 National Geospatial URI-system in Finland

The National Land Survey of Finland (NLS) hosts a centralised national spatial data URI redirection service. Both INSPIRE spatial data and other spatial data will be given URIs there. From the service, the queries are directed to data providers' own service.

The recommended structure for URIs of INSPIRE-harmonised data (in JHS 193): http://paikkatiedot.fi/so/resourceIdentifier/theme/class/localid/(versionid)

In future, queries from the centralised national URI service can be redirected to Statistics Finland´s service and from there, data can be accessed in many formats (e.g. GML and RDF/XML).

3.2.4 Results of the first pilot testing unique identifiers and RDF for geospatial data at Statistics Finland

As part of the geospatial project "Renewing of GIS technology", a pilot for testing how to create unique identifiers for geospatial data and how to use them in providing linked open data was carried out at Statistics Finland in 2016. Municipalities were chosen as the geospatial data to be tested in this pilot.

In the pilot, a functioning technical infrastructure for a redirection service was created. Technical solutions about how to provide geospatial data as linked open data were tested as well. In addition to this, ways to create unique identifiers for geospatial data were established and first ideas about how the identifiers of spatial objects could be combined with the most recent areal classifications were tried out.

The results and experiences of this pilot can be utilized in future work integrating areal classifications and corresponding geographies.

3.3 Plans to integrate areal classifications and corresponding geographies in future

Although geospatial information plays a vital role in the statistical production process, the geographies have not been connected to the corresponding areal classifications used in statistical datasets yet. We are planning to complete the work done in the first pilot and create linkages between areal classifications and corresponding geographies in a development project starting in 2018.

The main goal of the project is to create links between areal classifications and corresponding geographies so that they can be used together throughout the statistical production process, published as linked open data (LOD) and used to create new interactive map applications. The project consists of several specific targets that each build and improve on the previous one.

The first specific target is to implement an identifier system to manage Uniform Resource Identifiers (URIs) for areal classifications and geographies. Unique identifiers will enable creating links between objects both within Statistics Finland and shared between organisations.

The second specific target is to create ontologies for integrating areal classifications and geographies, using the URIs created and taking into account GSIM and INSPIRE. These ontologies will define and document the relationships between the different objects and serve as a base for practical information models.

The third specific aim is to implement a practical RDF-based (Resource Definition 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 number of possible applications. There will also be a cross-organisational pilot with the National Land Survey of Finland where the solution will be utilized in the geospatial data production process.

In this work, we can utilize the results and experiences of the first pilot tests already done in 2016. According to these experiences, careful planning is needed to combine two different data formats and many different standards. Creating well-defined ontologies is also a needed core requirement. A major challenge we are facing is that several versions and variants of areal classifications exist in our production as we do not reuse Classification Item elements. Some minor challenges have already been won, for instance, Scandinavian letters in identifiers are to be avoided already in the primary systems.

As an outcome, this project will offer ontologies, data models, linked data and practical applications using these linked data, which will be distributed freely.

3.4 Conclusion Case B

When areal classifications and corresponding geographies are linked, it is possible to use these together during the whole statistical production process. In future work to create these linkages our aim is to use the GSIM to describe the geospatial objects and linkages to them. According to preliminary planning and pilot testing, this approach could be beneficial for statistical production.

4 Conclusion

The GSBPM and GSIM are generic standards for statistical production. Nevertheless, geospatial data and processes are not described within them.

However, these models, the GSBPM and GSIM, can be utilised for describing geospatial objects and processes in statistical production. This requires that the information objects and process phases into which these geospatial objects are linked are identified. It is also crucial to communicate the results and ways to interpret these standards in connection with geospatial information to develop the statistical standards and their guidelines further.

Further reading

Generic Statistical Business Process Model GSBPM v5.0. UNECE, 2013.
https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

Generic Statistical Information model GSIM v1.1. UNECE, 2013.
https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model

GSIM Statistical Classification Model. UNECE, 2013.
https://statswiki.unece.org/display/gsim/Statistical+Classification+Model

Guidelines for the encoding of spatial data.
http://inspire.ec.europa.eu/documents/Data_Specifications/D2.7_v3.3.pdf

Implementing ModernStats Standards, Linked Open Metadata Design guidelines, v1.0. UNECE 2016.
https://statswiki.unece.org/pages/viewpage.action?pageId=129172661&preview=/129172661/129335617/HLG-MOS%20-%20IMS%20Design%20Guidelines_Jan2017.docx

INSPIRE Generic Conceptual Model.
http://inspire.ec.europa.eu/documents/Data_Specifications/D2.5_v3.4.pdf

JHS 193 Unique identifiers of geographic data. Recommendation.
http://docs.jhs-suositukset.fi/jhs-suositukset/JHS193_en/JHS193_en.html

JHS 193 Unique identifiers of geographic data. Annexes. *http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/193*

Kaukonen Essi and Leino Jennika. Implementing the GSIM Statistical Classification Model – the Finnish Way. Workshop on International Collaboration for Standards-Based Modernisation, 2015.
https://statswiki.unece.org/display/Standards/Geneva%2C+5-7+May+2015?preview=/112133421/113147945/Topic%20II%20-%20Finland%20-%20Paper.pdf

Linked Data. Wikipedia.
https://en.wikipedia.org/wiki/Linked_data

Ontologies. W3C website.
http://www.w3.org/standards/semanticweb/ontology

Radini Roberta, Scannapieco Monica and Garofalo Giuseppe. The Italian Integrated System of Statistical Registers: On the Design of an Ontology-based Data Integration Architecture. New Techniques and Technologies for Statistics, 2017.

RDF 1.1 Concepts and Abstract Syntax. W3C recommendation.
http://www.w3.org/TR/rdf11-concepts/

Stoop Ineke. Users of statistics, Who are they and what do they need/want? ESAC Meeting Helsinki 12.6.2017.
http://tilastokeskus.fi/ajk/tapahtumia_en.html/statistics-when-facts-count

Tilastokeskuksen strategia 2016-2019 (Strategy of Statistics Finland 2016-2019). Tilastokeskus 2016.
http://www.stat.fi/static/media/uploads/org/tilastokeskus/strategia_2016-2019.pdf