Statistics Finland

UNECE ModernStats World Workshop

Theme (i) Where to begin with standards based modernisation

GSIM implementation at Statistics Finland

Kaukonen Essi, Daniel Davis, Leino Jennika, Pasila Aura, Räikkönen Toni,
Saloila Mikko, Saranpää Tuukka. firstname.lastname@stat.fi

ABSTRACT: Statistics Finland has not had a profound conceptual model for the statistical production and, therefore, the GSIM-model has been welcomed. The first studies of the GSIM-model and translation experiments were carried out around 2015 by few experts leading yet to no true implementation efforts. The interest arose after a physical implementation of the GSIM Statistical Classification Model into our renewed Classification System launched in production 2016. After that the GSIM Model has been utilized in different ways in a couple of international and national cooperation projects providing a common language, basis for local implementations in research area and a tool for integrating statistical and geospatial approaches. At the same time the Information Architecture at Statistics Finland has developed further including some core elements of the GSIM model which have been introduced to our top managers. A couple of internal development projects are already using GSIM as the framework in different parts of our organisation. The risk of slightly different interpretations of the GSIM model in these projects is evident and, therefore, the solution could be to create a coherent complete GSIM-based information model fitting into our environment. While creating it also the national architectural guidance and tools for semantic interoperability as well as the changing needs of our customers and in-house processes should be taken into account, not to forget a working governance model.

First steps in implementing GSIM around 2015

From the official adoption of GSBPM to first experiments on GSIM

Finland has gone the typical way of adopting the UNECE standards for modernisation. The Generic Statistical Business Process Model (GSBPM) was introduced first, after that the Generic Statistical Information Model (GSIM) and the Common Statistical Production Architecture (CSPA) have also been in focus, nevertheless, not yet gaining as much effort as GSBPM.

The GSBPM was officially adopted as the framework for statistical production in 2015 as such; no national changes except for Finnish translation have been made to this model. GSBPM is included in our 2016 to 2019 strategy as the reference model for our core business process. Translation of the GSBPM was introduced also to the other producers of official statistics in Finland and they are available on our website. In practice, the GSBPM has been used in development projects as the process model framework leading to more coherent processes, mainly by the support of our IT Architect group. GSBPM is covered in the internal education program. Intranet pages supporting the use of GSBPM exist.

During the time of the official adoption of GSBPM, the GSIM model was explored for the first time. First attempts to utilize GSIM were made in Eurostat ESRBs –project. Inside Statistics Finland a couple of workshops to study the model and to translate it into Finnish were held where experts from Standards and Methods Unit and IT division took part. The translation was not finished as the potential of GSIM was not yet understood.

At that time, no commonly used vocabulary or conceptual model for statistical production existed at Statistics Finland. The CoSSI data model developed at Statistics Finland and published initially in 2002 has been used extensively in describing datasets and also other components in dissemination. CoSSI consists of several different modules for predefined concepts, for example classifications, datasets, quality metadata, and publications, and in addition to these, separate metadata sections that can be attached to any of the concepts mentioned. The idea of the CoSSI model was to create a generic model that could be used for purposes of different kind. However, it turned out to be somewhat too generic to meet some of the requirements of statistical production of today. Hence, expectations

towards GSIM have been increasing during the past few years. Because of the modular approach it is possible to continue using CoSSI in some areas while introducing GSIM and GSIM-related physical data models in others.

GSIM Statistical Classification Model and the new Classification System

Around 2015, Statistics Finland made the decision of adopting the GSIM Statistical Classification Model for the new Classification System. We did not implement the whole Node Set structure of the GSIM v.1.1 as such due to historical, practical and timing reasons. (See also Kaukonen and Leino 2015, Kaukonen et al 2017)

Before the first physical implementation of GSIM, the model was studied closely and the implementation was designed by a group of classification and IT specialists. International cooperation was also undertaken especially with the Norwegian statistical office. In addition, detailed tests with some individual classifications were performed. The GSIM model was also compared with the CoSSI model for disseminating classifications and the classification schema that was in use in the existing classification system.

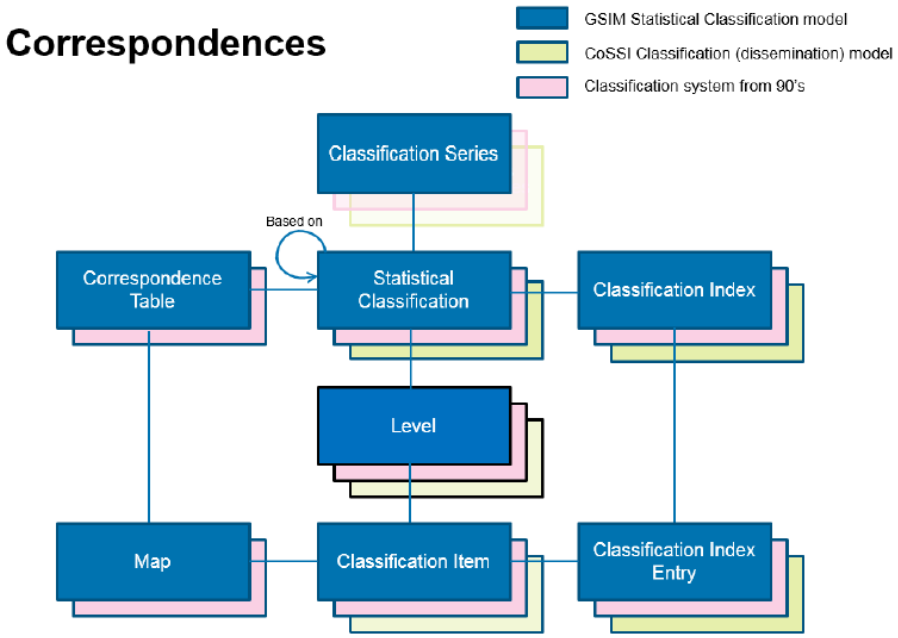


Figure 1. Comparing GSIM Statistical Classification Model to models already in use at Statistics Finland

Implementing the GSIM Statistical Classification model was a success, which is not a surprise as it is widely used in statistical organisations. It was worthwhile to take the time to explore and test the GSIM model for that way we got to know it well and were able to make it as conceptual and logical base in our system. The GSIM model fitted mostly in our purposes just as it was and we made only a few extensions to it. (See also Kaukonen and Leino 2015, Kaukonen et al 2017)

All in all, GSIM provides us today a sounder terminology in this area. The terminology in Finnish language used in our organisation is now fixed and this has made communication within our organisation easier. In addition, GSIM facilitates discussions about classifications with colleagues in other NSO's.

GSIM in cooperation projects with other organisations

NordMan-project

NordMAN (Nordic Microdata Access Network) was a cooperation project between the Nordic NSI's funded by NordForsk. The overarching aim of the NordMAN project was to improve the service provided by the Nordic NSI's for researchers using cross-nordic microdata.

One objective of this project was to work towards a solution of harmonized metadata for microdata hosted by the Nordic NSI's. One of the more pressing needs of researchers when using data from different sources (e.g. different NSI's) is to know to what extent the data are comparable and available. At an early stage, the project group reached consensus on using the international standard GSIM as a translator and bridge between different documentation and metadata systems in our respective countries. (NordMAN-project, 2016)

In the project it became evident that there is a need to go around the limitation of just one Unit Type object per Variable. The solution that was developed in the project is pictured below

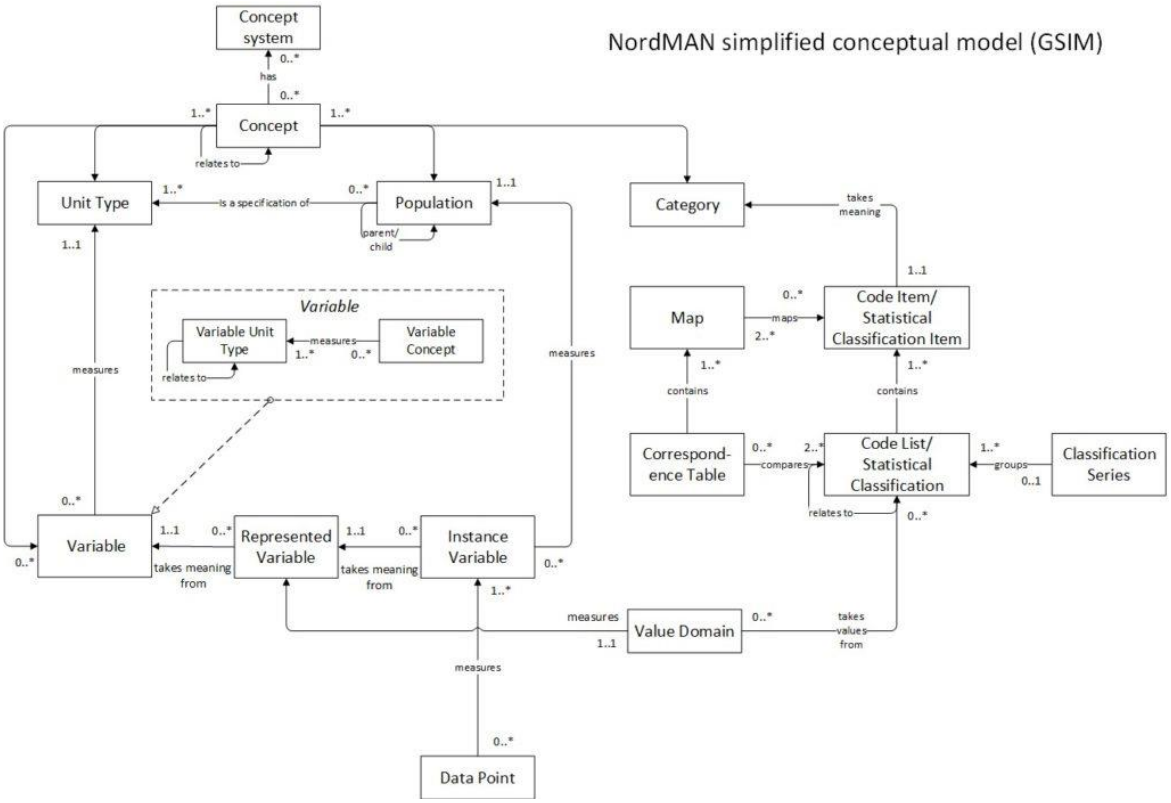


Figure 2. NordMan simplified conceptual Model (GSIM)

This more detailed conception of the Variable object enables to model more elaborate Variables like “Mother’s country of birth”. The need to attach more than one Unit Type to a Variable was then forwarded to GSIM revision group.

The NordMAN project offered a wider platform to discuss and share understanding over GSIM. This enabled Statistics Finland to advance with the implementation of GSIM also inside the organization. This will eventually make data more compatible cross Nordic and further.

Isaacus-project

Standards were also strongly implemented in national project called ISAACUS. The project aim was to develop national metadata description model and database for digital information resources, with a special focus on health data. This project was funded by the Finnish Innovation Fund Sitra. The project was led by National Institute for Health and Welfare. Statistics Finland and Finnish Social Science Data Archive (FSD) acted as partners.

GSIM was chosen as the standard base for the implementation of the data description system. From these experiences we found out, that the basic structure of the Concept part of GSIM is also more than enough to describe health data and so called big data in data lakes.

National Institute for Health and Welfare is also a producer of Official Statistics of Finland. This creates a natural link to international standards for statistical community. GSBPM was discussed in the project before diving in too deep into GSIM. This approach eased the perception of GSIM when the business process was outlined first. Experts in the National Institute for Health and Welfare also felt that the business process is something that needs to be implemented first before the benefits of GSIM can be experienced on an organizational level. It is also important to keep in mind that the National Institute for Health and Welfare has also other business processes than just the process for producing statistics.

This project gave Statistics Finland a lot of new information about national Statistics production. Especially how GSBPM and GSIM can be used to translate and possibly harmonize the ways we produce different statistics on a national level. The project gave a lot of insight how beneficial this can be to the researcher, statistician or other end users of data.

IGALOD-project

The geospatial approach has not yet been clearly visible in either GSBPM or GSIM. In the UNECE workshop for integrating statistical and geospatial standards November 2017 the Australian proposition how to add the geospatial approach into GSIM was introduced (see Walter and Brady). This solution is pretty much the same as the practical actions to integrate areal (geospatial) classifications and their respective geographies in the Integration of Geographies and Areal classifications as Linked Open Data (IGALOD) -project. This project for the years 2018-2019 will be carried out together with Finnish Land Survey Institute. Clearly, the main challenge in the project is to combine the statistical and geospatial worlds, their concepts and actual datafiles. In practice, this will be done by using ontologies and RDF solutions.

Understanding the role and coverage of GSIM

A major breakthrough in understanding the role and coverage of GSIM was achieved during the international projects and the Information Architecture project of Stiina-program (described in more detail further below under heading Stiina). A clear distinction between Domain Concepts and GSIM Concepts was established, as illustrated in the picture below. This distinction is vital to understand whenever you discuss the information assets of your organisation.

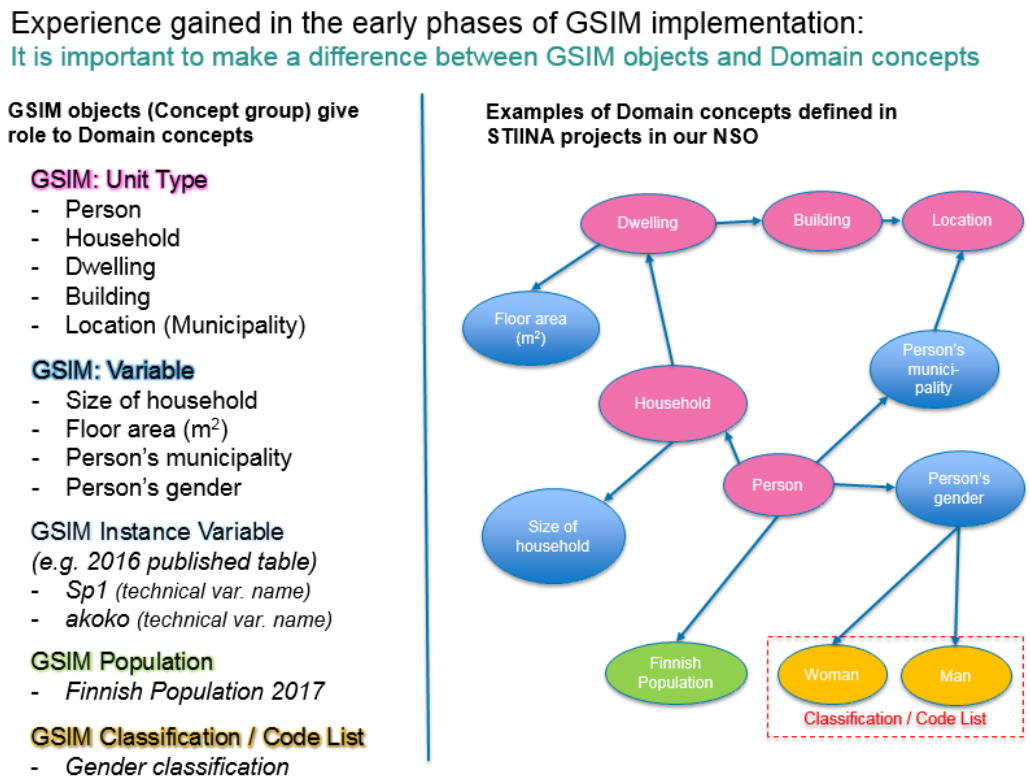


Figure 3. It is important to make a difference between GSIM objects and Domain concepts and the relationship between them.

GSIM in internal projects 2017 and 2018

During the past year and this year the use of GSIM in internal projects has been rapidly increasing. Information Architecture work, plans for a common Identifier Service and the GSIM-based metadata sandbox of Stiina-program are described below. In addition to these, partly GSIM-based solution for creating metadata for statistical tables is currently in pilot phase. Interest in utilizing GSIM is also rising i.e. in connection with developing the process governance system further and modernizing the data collection as well as structuring disseminated data and metadata in general.

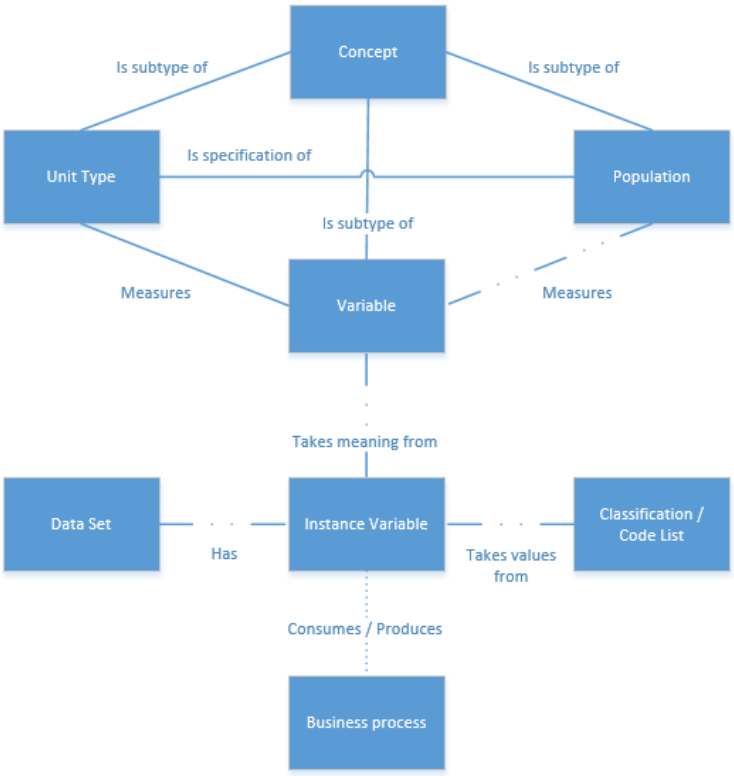
Information Architecture and communicating GSIM to the general audience

In 2017, a program for the development of Information Architecture for Statistics Production was launched in Statistics Finland. The objective of the program is to coordinate and facilitate cooperation and interoperability between the various development projects underway in the organization. The program has also organized several sub-projects to produce top-level models and descriptions to be used in the more specific development work. One

additional goal of the program is also to follow and participate in the information architecture development in the Finnish public administration.

GSIM is not as easy to communicate to the general audience, statisticians and management as is GSBPM. One reason for this is the general abstract nature of information and structures of information, another is the number of detail in the GSIM model. The first sub-project of the Information Architecture program aimed to tackle this issue in 2017. To inform the managers and statisticians at Statistics Finland a core model was created where only 8 core elements of GSIM, translated to Finnish, were selected (see the picture below). The Board of Directors accepted these core elements as a first version of the reference model for the information serving the statistical production process. The model will be extended in the future.

The other purpose of this action was to give those development projects which had already started utilising GSIM a clear sign that they can move on with GSIM and that these core elements of GSIM are the ones they can rely on in their solutions.



Picture 4. GSIM core elements 2017 with examples

Developing Social Statistics Integrated Information System (STIINA) following GSIM and other standards

STIINA (Social Statistics Integrated Information System) is a wide program including several projects aiming at modernising the production of the social statistics. The work is based on international standards, including GSIM. The original focus of the project was to renew production processes and create a new data warehouse, but it was soon realised that wider modernisation was needed. It was not possible to build new type of services for renewed processes and still rely on the current, relatively old metadata solutions and the CoSSI model used in many of them.

Thus, as a part of a Proof of Concept -project (named Stiina kok2) focusing on data collecting, a set of GSIM-based metadata elements were created and added into a “sandbox” for practical testing. The elements are presented in the following figure. They are mainly from the GSIM Concepts group, some process metadata elements also from GSIM Business group. In the sandbox, these reusable objects are used to store descriptive metadata, technical data and processing rules. The new services created in the project are metadata-driven and they utilize these elements in the processing.

The solution will be further developed and extended during the rest of the year 2018 in a project focusing on metadata solutions.

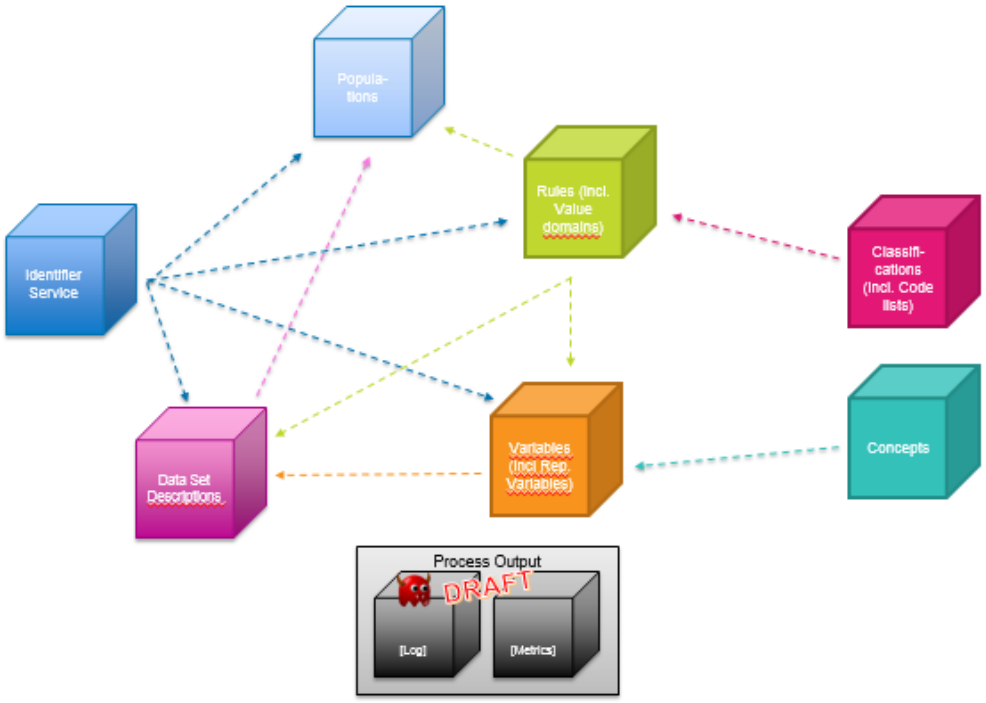


Figure 5. Social Statistics Integrated Information System and simplified view to the GSIM based Metadata "sandbox"

Future plans: Creating an identifier service for statistical production as well as a GSIM-based national Information Model

Winter 2017-2018 a new project targeted in creating a governance model and a roadmap for an Identifier Service for objects used in the statistical production. The contents of the Identifier System i.e. to what objects it will deliver identifiers, will be added to the system after first adding them to a national GSIM-based model, see the Figure below. During the project, cooperation with a national group working with identifiers, the National Library of Finland and National Land Survey of Finland was successful, not to forget fruitful discussions with our Australian colleagues.

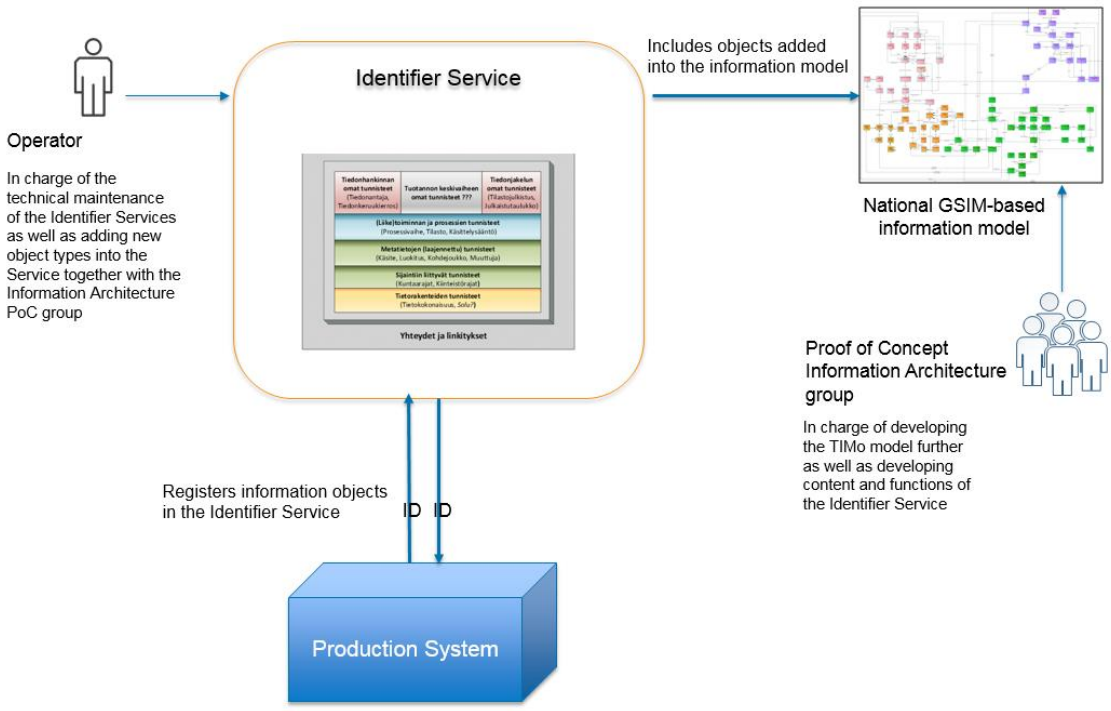


Figure 6. Governance Model of the Identifier Service, future.

Challenges and solutions today on the way to corporate-wide GSIM implementation

During the past two years, we have made a clear jump from pre-implementation phase to the early implementation of GSIM framework. Following ModernStats survey 2018, our state today could be described as "Use of the GSIM (or the national version of it) is spreading, but limited to individual projects in an isolated manner".

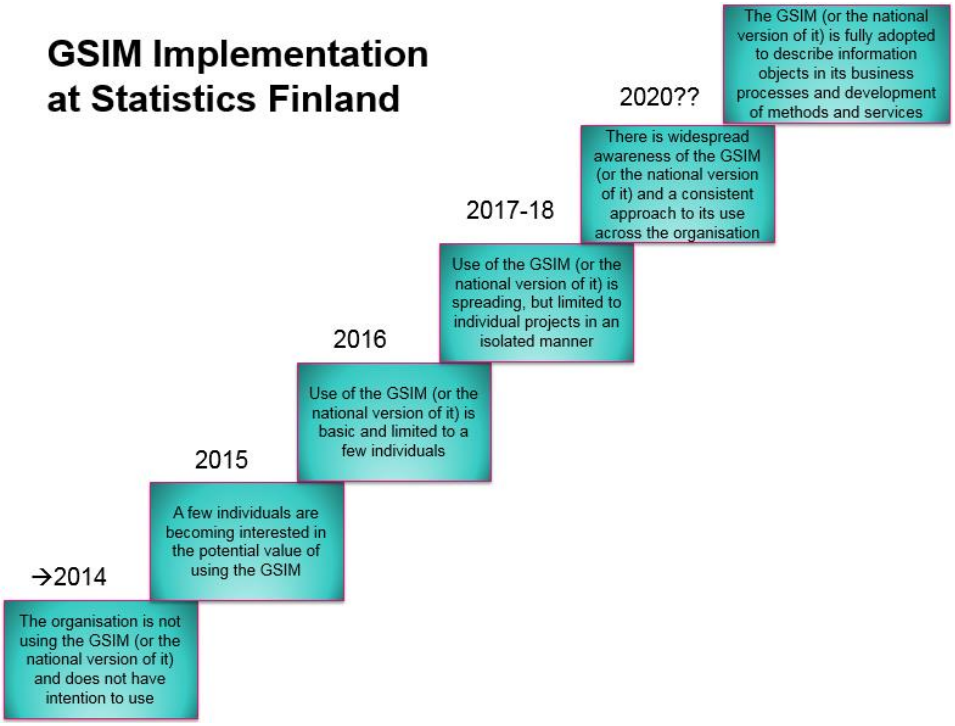


Figure 7. GSIM Implementation at Statistics Finland

The imminent risk in this state is, that the organisation ends in creating several differing implementations of GSIM and thus, leading to no real gains. To avoid this, coordination is needed at corporate level.

We can aim at creating a consistent approach to use GSIM across organisation by creating a national adaption of GSIM to suit our needs. Hence, our plan is to create a national information model for statistical production based on GSIM. This model could, in practice, be created one step at a time taking into account the needs and propositions of chosen development projects.

To govern this work a Proof of Concept Information Architecture group is proposed. Currently, we do not have IA or BA groups in our NSO or experts named specifically for these actions. The main tasks of the group would be to develop our national GSIM-based model further. This group would also be in charge of the development and content governance of the new Identifier Service.

By creating a full national GSIM-based model, we can have, in the future, reusable identified information for the whole statistical production serving also the needs of our customers. Furthermore, it would be a strong basis for software development following CSPA principles.

What else would we need to take the next step to corporate-wide GSIM implementation? Naturally, the support for the chosen path from upper management is necessity. In practice, supporting material and training sets in Finnish language needs to be created. Last, but not least, continuous cooperation and benchmarking with other NSO's could boost the development in a way that we can, in fact, find us standing on the next step of GSIM implementation ladders during the year 2020.

Further reading and useful links

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

Isaacus-project, National Metadata Descriptions 2016-2018.
https://thl.fi/en/web/thlfi-en/research-and-expertwork/projects-and-programmes/national-metadata-descriptions

Kaukonen Essi and Leino Jennika. Implementing the GSIM Statistical Classification Model – the Finnish Way. UNECE 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

Kaukonen Essi, Tammisto Rina, Pihlajamaa Tuuli Pihlajamaa, Davis Daniel, Leino Jennika.
Towards connecting geospatial information and statistical standards in statistical production: two cases from Statistics Finland. Joint UNECE/UN-GGIM Workshop on Integrating Geospatial and Statistical Standards 2017.
https://statswiki.unece.org/pages/viewpage.action?pageId=151453739&preview=/129178137/141951145/Towards%20connecting%20geospatial%20information%20and%20statistical%20standards_Statistics%20Finland.docx

Modernisation Maturity Model. UNECE, 2017.
https://statswiki.unece.org/display/RMIMS/Introduction+to+the+Modernisation+Maturity+Model+and+its+Roadmap

Nicholas Nick, Ward Nick, Blinco Keery. A Policy Checklist for Enabling Persistence of Identifiers. D-Lib Magazine 2009.
http://www.dlib.org/dlib/january09/nicholas/01nicholas.html

NordMAN-project, A cooperation between the Nordic NSI's funded by NordForsk, 2015-2017
http://nordman.network/

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

Walter Tom, Brady Martin. Location information in statistical modernisation transformation, paper and presentation. Joint UNECE/UN-GGIM Workshop on Integrating Geospatial and Statistical Standards 2017.
https://statswiki.unece.org/pages/viewpage.action?pageId=151453739