**Automatized Validation and Pseudonymization of Data Sets
– New Data Acquisition Process**

Aura Pasila, Senior Statistician, Statistics Finland, aura.pasila@stat.fi

Lea Lof, Senior Statistician, Statistics Finland, lea.lof@stat.fi

Noora Seppä, Senior Statistician, Statistics Finland, noora.seppa@stat.fi

Toni Räikkönen, Manager, IT Development, Statistics Finland, toni.raikkonen@stat.fi

**Abstract**

*New kinds of data sources change the ways data are collected and processed in statistical offices. The centralised data collection system of Statistics Finland is currently developed to tackle challenges such as complex data formats, daily or even hourly data deliveries, increasing number of incoming files and data protection. The aim of the development work is to automatically validate and pseudonymize all incoming data sets irrespective of their original format, source or type, in order to provide statistical units with data of uniform quality. The new process is built on small independently operating microservices, a process control system and a GSIM-based (Generic Statistical Information Model) metadata system. Microservices are metadata-driven and automatically operated by the process management system. Metrics produced by the services can be used to monitor the process and the quality of the data sets. At the moment, the new process is in the implementation phase. The expectation is that when it is in use, the increasing number of data files can be acquired and consistently validated with current resources.*

***Keywords:*** *Data collection, automatization, data validation, pseudonymization, data quality*

**1. Introduction**

Statistics Finland has been using administrative data and other secondary data sources in statistics production for decades. An important milestone was achieved in 1990 when the first register-based population census was completed. Since then, the use of different secondary data sources in statistical production has increased continuously. Not only have new sources been emerging, but simultaneously, many traditional data source keepers are modernising their databases technically. In addition, new techniques such as web scraping, enable novel ways to collect data.

Today, Statistics Finland receives data from approximately 200 data set sources per year and the total number of incoming files is counted in tens of thousands, survey data sets not included[[1]](#footnote-1). The data sources are changing rapidly. Data from the data source keepers become available faster and more frequently than before, but on the other hand, more work is needed to compile data sets so that they can be used for statistical purposes as the quality and technical structures of the new sources often differ from the previous ones. Also, self-service is often needed to get hold of data instead of ready-made data files delivered by the data source keeper.

At Statistics Finland, the solution to these challenges has been to centralise data collection tasks, data acquisition process and expertise needed to perform them. In practice, the challenges mentioned above have created the need to enhance data management, make processing of incoming files more effective, remove overlaps and find general quality standards across different data sets.

**2. History of the centralised data acquisition process**

2*.1. Towards centralised data collection*

Before 2013, each unit producing statistics usually acquired the necessary data files themselves. Due to organisational changes and Statistics Finland’s Information Communication Technology Strategy for 2015–2019, a new project called Havas was set in April 2013 to centralise the acquisition of administrative data from various data source keepers (Laurila, Eskelinen and Sisto, 2016, pp. 1–2). In 2015, Statistics Finland began receiving data through the new data acquisition process operated by the data unit. The new system proved to be safer, quicker and more reliable than the previous decentralized situation.

*2.2. Description of the current data acquisition process*

The current centralised data acquisition process version 1.0 works as follows:

* The data source keeper sends the files to Statistics Finland’s secured server based on a common agreement describing the timetable and content of deliveries.
* At Statistics Finland the received file is identified by the file name and the sender’s login-ID and directed to the right target directory. If the data file is not identified, an e-mail notification will be sent to the team of system owners.
* Flat file data files can be transformed manually into SAS file format with an application that uses pre-saved metadata information of the data file. This process also validates the file against metadata and prints validation results into an HTML report which is delivered to unit receiving the file.

The centralised data system has now been in production use for several years. Statistics Finland has been able to standardise its processes to internal stakeholders and data source keepers and to reduce overlapping work within Statistics Finland. However, there has existed development needs in the current system since the beginning:

* Level of automation: the process includes manual steps.
* New data sources: the process is only suitable for technical check-up and validation of simple flat files.
* Controlling the process: the user cannot control or manage the whole process.
* Reporting: the current system is based on e-mail notifications, which makes it difficult to gather information from the whole system (e.g. the total amount of data sets received through the system).

As the number of new data sources keeps increasing, the need to develop the current system has become more urgent. There are also still some files that for various reasons have been going directly to units producing statistics without centralised acquisition. In the future, these files should be acquired through the centralised process.

*2.3. Developing the data acquisition process and statistical production*

Many of the development needs in the data acquisition process mentioned previously were also pointed out as a wide social statistics development program called STIINA (Social Statistics Integrated Information Architecture) was planned (Nieminen, 2015). It seemed that in order to modernise the statistical processing phase, the data collection and the existing acquisition process also required renewal. Despite the implementation of the first acquisition process, there still existed overlapping data sources, different kinds of validation procedures, and a lack of information on the acquisition process. There was also a wish to develop future-proof solutions for storing and using new kind of data, such as the real-time Income Register coming into use in 2020. An additional requirement, especially related to social statistics, was set by the GDPR legislation (General Data Protection Regulation): the process should universally take care of the pseudonymization of incoming data meaning removing names and replacing social security numbers with a pseudonymized IDs.

The vision was to get incoming data to the statistical units as quickly as possible to accelerate the whole production process. All data sets were planned to be validated and processed in a standardised way irrespective of the original format to guarantee a certain quality level. In order to create as general a process as possible, the process was to be built on the principles of the enterprise architecture method (JUHTA, 2018) and international models and standards, mainly GSBPM (Generic Statistical Business Process Model), GSIM (Generic Statistical Information Model) and CSPA (Common Statistical Production Architecture) (UNECE, 2019a; UNECE, 2019b; UNECE, 2019c). The following development and implementation work in 2017–2019 was done in collaboration between the STIINA program and the Data Collection Department.

**3. Planning the data acquisition process 2.0**

*3.1. Requirements for the new process*

As the planning of the new acquisition process began, the practical requirements that the process must fulfil were specified on a more detailed level. First and foremost, the process should be fully automated from the moment a new file is recognised to have arrived, to the point when it will be stored into the raw data repository. Human intervention should be required only when detecting erroneous data. The user should be able to visually monitor all the phases of the process and metrics of all operations must be available for detailed inspection purposes.

Secondly, the process must be completely metadata-driven; no operations of the data sets are carried out without the corresponding metadata. This way, it is known exactly, what kind of data has arrived at Statistics Finland. Also, metadata is needed to control the services that are part of the acquisition process. The process consists of a number of independent microservices that each perform a specific task to the data files using the GSIM-based metadata descriptions.

The plan was to divide the processing of files into different zones that each perform certain tasks. The zones are the entry zone, the structural validation zone, the pseudonymization zone, the analysis zone and the raw data repository zone (Figure 1). It should be noted that the pseudonymization is the only zone in which the data are modified, otherwise files are only validated and left untouched. In some rare cases, it might be necessary also to clean erroneous characters from the data.

The new data acquisition process must be able to store and process most typical open file formats (CSV, XML, JSON). Thus, no transformation of the files is required before loading the data into the production databases. The whole acquisition process ends in the raw data zone where files are left in their original format to be fetched by the statistics.

**Figure 1. Data Acquisition Process 2.0**



Lastly, the new process must be able to process all different types of data acquisition methods. These include, but are not limited to, surveys, administrative registers, web-based questionnaires and transfers from modern application programming interfaces (APIs).

*3.2. Technical planning principles and vision*

As with most new information systems at Statistics Finland, the new data acquisition process is also implemented using a microservices architecture. The microservices mostly use the by-reference data access model, thus fetching the data files using a pointer to the storage location. Currently, a file-system-based storage solution is in use. This implementation model follows the CSPA quite nicely (UNECE, 2019b).

The microservices-based approach has several advantages. Each service can be developed and deployed independently of the others. A service is only responsible for performing a very limited task and is executed in an isolated environment, meaning a malfunction of the service does not affect the other services in the process. In addition, new services can be introduced to the process easily and deprecated ones removed as well. Some of the microservices can also be used outside of the data collection process phase.

The microservices require an orchestration platform that is responsible for executing the services in the correct order and reacting to the response messages of each service. The in-house process management system is used to orchestrate the data acquisition process. This allows modelling the process with the BPMN (Business Process Model and Notation) graphical representation and connecting the necessary services to the process tasks. The process can then be instantiated and visually monitored while it is executed. If an error takes place in the execution, the process will be halted, and an expert must resolve the erroneous task. Once the problem has been resolved, execution can be continued, restarted or exited.

*3.3. Using metadata to control the process*

Metadata-driven services were seen as the key to create an automatized and controllable process. However, it seemed too difficult, or in some cases impossible, to build services that would use the current Variable Editor system developed at Statistics Finland storing metadata related to data sets. The system is based on a CoSSI model (Common Structure of Statistical Information) (Statistics Finland, 2008). The main difficulty with the Variable Editor is that it can store only very simple flat file descriptions and most of the new data structures are technically much more complicated. In addition, saved metadata objects are not reusable (Kaukonen and Saloila, 2016). Also, the service does not include the possibility to attach processing rules to variables.

The solution was found from the emerging GSIM model. There had been previous Nordic development work to build a common GSIM-based metadata model for creating data file descriptions (Kaukonen and Saloila, 2016). Also, Statistics Finland’s Classification Editor was already renewed along the lines set by the GSIM model. From the viewpoint of data acquisition, the business group of the GSIM model includes rule and process elements that were considered very useful. One of the cornerstones of the new process is that all incoming data files and the processing rules must be described into the metadata system, otherwise the file does not move through the acquisition process.

**4. Creating a new metadata system, microservices and database for the acquisition process 2.0**

*4.1. A new metadata system*

A new GSIM-based metadata system named Metsy was planned to support the microservices. The key elements of the new system are represented variables, rules and unit data structure descriptions (Figure 2). The represented variables (such as *age* or *name*) are stored to a variable storage from where they are attached to the unit data structures to create an exact description of a data file. When represented variables have been linked to the unit data structure description, rules can be attached to them. An example of a rule could be a pseudonymization rule indicating that the specified variable in the unit data structure should be pseudonymized. The units producing statistics and data unit will create a unit data structure description including processing rules in cooperation when the data collection is planned.

**Figure 2. GSIM-based metadata elements used by metadata-driven services**



The new metadata system is more flexible than the previous one making it possible to describe even very complicated data structures by using sub-structures called logical records. All technical information related to a file which is needed by the microservices is attached to the unit data structure description (such as schema files). The idea is that the metadata-driven services will pick the right metadata description including the processing instructions based on a link between file-ID and metadata description ID. The link must be specified in advance.

*4.2. Service catalogue*

A selection of independent microservices has been built to validate and analyse the acquired data. According to the microservice principles, each service is responsible only for its own specific task. Services are metadata-driven as they reflect whether the acquired data matches up to the metadata description. All the services can process data in a few different open file formats. For the time being, the supported formats are CSV, XML, JSON and Microsoft Open XML documents.

Currently, the following services are implemented in the new data acquisition process:

* **Schema Validation Service**: Validates the data file against the corresponding schema (JSON, XML, CSV).
* **Rule Validation Service**: Validates the data file against the unit data structure description, validations include data type and other structural checks.
* **Range Validation Service**: Validates variable values in files against the given numeric range.
* **Regular Expression Validation Service**: Validates variable values in files against the given regular expression.
* **Duplicate Validation Service**: Finds any duplicate values from defined variables in files.
* **Frequency Analyser Service**: Calculates frequencies from the data file.
* **Classification Validation Service**: Validates variable values in files against an enumerated value domain (i.e. classification, code list).
* **Xpath Validation Service**: Executes XPath statements against a given XML document.
* **Janitor Service**: Clears the nonprinting characters from the data file, for example unnecessary empty spaces.
* **Pseudonymization Service**: The personal identification number is replaced with a unique pseudonymized identifier. The same PIN gets the same pseudonymized identifier in every dataset. Links between PINs and pseudonymized identifiers are stored in a dedicated database. First names and last names are stored in the same database as the pseudonymized identifiers and the name fields are cleared from the data file. E-mail addresses are also cleared from the data file and stored in the database with names and PINs. Pseudonymization doesn’t change the structure on the datasets. The pseudonymized identifier is completely random, and it doesn’t include any information about the person – compared to a PIN that can be used to determine a person’s date of birth and gender.
* **Xml Signature Validation Service**: The service checks that the signature in the XML file is valid.

Pseudonymization Service and Janitor Service write edited data as a result, but all the other services do not edit the data. In addition, all the services produce information about their observations. A separate Metrics Service saves the results in a dedicated Metrics database.

*4.3. Database for process controlling information*

In addition to unit data structure descriptions that are used to control the services, process controlling metadata is needed to create an automatized, metadata-driven process. The solution is also based on the GSIM model, which was used to outline concepts and their relationships, and to identify the necessary data objects.

The database was built based on the conceptual model presented in Figure 3. A simple application was created to enable browsing and updating of the database. However, this is only a temporary solution to store and maintain the information needed for the new data acquisition process and the content covers only the minimum information to run the process. In the future, this data storage will be replaced with an extensive database and a more advanced application.

**Figure 3. Data acquisition and GSIM conceptual model**



*4.4. Processes in the process management system*

The first stage of process planning is to identify the acquired data. Once the file is recognised, an identifier for the corresponding process can be fetched from the database and the right process can be started in the process management system. The processes have parameters that control which services are to be executed for each data file. The parameters are defined in the process controlling information database and passed to the process by the same program that identifies the data.

The process management system is used to perform commands to the services (Figure 4). Each service must have a service attribute defined in the process management system so that the process management system can call the service. The process management system and the microservices-based architecture enable modifying processes and creating and adding new services quite effortlessly.

**Figure 4. An example of the acquisition process in the process management system**



The services report the errors found during the process. To manage errors, different error handling routines can be defined in the process management system. Usually the errors have to be studied manually.

**5. Initial results from the implementation**

The proof-of-concept process was created in autumn 2018. The file placed into the input file was automatically delivered to the output file through schema and rule validation, pseudonymization, frequency analyser and classification validation service. The results were saved to a database (Figure 5). However, the process was still demonstrated in the development environment and had to be started manually.

**Figure 5. An example of the result metrics report created by Microsoft Power BI**



Since autumn 2018, the process has been developed further. In short, testing and production environments for services are in use, a wider selection of services has been added to the acquisition process and the process can be controlled by saving information to a database. There are still some things under development before the first file can be processed in the production environment, most importantly, the services have to be updated to use the newest metadata system version. This work is still in progress.

**6. Conclusions**

As the process is now in the implementation phase, it is still too early to estimate the outcome of the renewal. Based on the proof-of-concept phase, it can be safely stated that the acquisition process can be automatized, is able to handle complex data structures that the old system was not able to process and perform new tasks that the previous version was not able to perform, such as pseudonymization.

The automatization is expected to curb the amount of work needed to manage the data acquisition process, even though the number of new data files is increasing, e.g. multiple data files from the Income Register are expected to be delivered daily from 2020 onwards. The plan is to gradually expand the new process to practically acquire all data sets entering Statistics Finland. The aim is also to ensure that same validations and other processing tasks are performed also for survey data, though in many cases, the data collection systems take care of these operations.

In terms of lessons learned so far, it has required a lot of work to carry out several renewals almost simultaneously, such as the creation of a pseudonymization service and a new metadata system. Both of them have required much more planning and work than was anticipated at the beginning. The development steps should be staggered carefully, and the previous phases should be finished before moving on.

The other important lesson is to find out whether there are some ready-made tools available that could be utilised. During the development work, a tailored portal for analyses was tested, but it was soon noticed that Microsoft Power BI and other ready-made analysing tools are more versatile and do not require that much in-house maintenance. In the future, it would be interesting to share solutions developed at Statistics Finland with other statistical officials to see if they could be utilised by others as well.

**7. References**

JUHTA (2018): *JHS 179 Enterprise architecture planning and development* [online]. Available at: <http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/179> (in Finnish) (Accessed: 15 May 2019)

Kaukonen E. and Saloila M. (2016): Käsitemalli GSIM ja Tilastokeskuksen fyysinen tietomalli CoSSI (*Concept model GSIM and Statistics Finland’s physical information model CoSSI*, in Finnish). Internal document, Statistics Finland. Unpublished.

Laurila S., Eskelinen J. and Sisto J. (2016): Reviewing the general and centralised system of collecting administrative data and how it responses to the guidelines of the Code of Practice and the Peer Review results in 2013 to 2015 [online]. In: *European Conference on Quality in Official Statistics*. Madrid: INE. Available at: <http://www.ine.es/q2016/docs/q2016Final00088.pdf> (Accessed: 12 April 2019)

Nieminen J. (2015): *Herttua uudistus ja laajennus -esiselvitys, loppuraportti* (*Final report on the preliminary study of the renewal and extension of Herttua,* in Finnish). Internal project report, Statistics Finland. Unpublished, available from Statistics Finland’s record office.

Statistics Finland (2008): *Common Structure of Statistical Information (CoSSI)* [online]. Available at: <https://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html> (Accessed: 28 June 2019)

UNECE (2019a): *Common Statistical Production Architecture* [online]. Available at: [https://statswiki.unece.org/display/CSPA/Common+Statistical+Production+Architecture](https://statswiki.unece.org/display/CSPA/Common%2BStatistical%2BProduction%2BArchitecture) (Accessed: 15 May 2019)

UNECE (2019b): *Generic Statistical Business Process Model* [online]. Available at: [https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model](https://statswiki.unece.org/display/GSBPM/Generic%2BStatistical%2BBusiness%2BProcess%2BModel) (Accessed: 15 May 2019)

UNECE (2019c): *Generic Statistical Information Model* [online]. Available at: <https://statswiki.unece.org/display/gsim/> (Accessed: 15 May 2019)

1. Information provided by Timo Järvinen, Statistics Finland. [↑](#footnote-ref-1)