**The role of administrative data manager in getting access to data and metadata**

Aivi Saar, Leading Methodologist, aivi.saar@stat.ee

Liisi Lillipuu, Leading Methodologist, liisi.lillipuu@stat.ee

**Abstract**

*According to Statistics Estonia’s strategy, our goal is to produce high quality statistics with as low administrative burden and as high efficiency as possible. In order to achieve this, we need to improve the use of administrative data and describe the related metadata in our metadata management system.*

*Statistics Estonia (SE) has centralised its administrative data management. The process goes through different departments and teams, where central administrative data managers work in different roles and process phases.*

*In 2017, the Statistics Design Department was created, where administrative data and metadata are managed centrally. The central role of administrative data managers in this department includes creating and developing partnerships with data owners, coordinating information exchange and consulting subject matter experts in Statistics Estonia, and managing data delivery contracts and data requests. The methodologists of the Statistics Design Department are also responsible for getting access to metadata of administrative data, negotiating with data owners and agreeing on the deadlines, data delivery formats and structures.*

*At the moment, SE uses over 100 different administrative data sources (state registries) in the statistical production process. Managing, describing and improving the related information and metadata of those sources is a challenging and ongoing process.*

**Keywords:** administrative data, metadata, data management

**1. Administrative data sources in Statistics Estonia**

The process of administrative data management is centralised in Statistics Estonia. It is considered to be best practice for NSIs to have a dedicated coordination unit for administrative data management. The result is enhanced visibility of the administrative data and increased efficiency of their use in the statistical production process (Eurostat, 2018).

The Statistics Design Department plays a central role in the process of managing administrative data and metadata. Metadata management was concentrated in the Methodology Department in 2004; managing and capturing administrative data was formerly the responsibility of the Data Warehouse Department. As SE has started using administrative data more and aims to create and develop closer partnerships with data owners, it was decided to concentrate the management of metadata and administrative data to the Metadata team of the Statistics Design Department. This centralised management enables to describe and harmonise the metadata of administrative data more efficiently.

Daily data capture is done through X-Road, but some datasets are sent to SE also by e- mail or FTP as encrypted .csv or .xls files.

X-Road is the backbone of e-Estonia: it is the data exchange layer that allows various public and private sector e-service information systems to link up and function in harmony. X-Road has developed into a tool that can write to multiple information systems, transmit large datasets and perform searches across several information systems simultaneously. Today, X-Road is implemented in Finland, Kyrgyzstan, Namibia, Faroe Islands, Iceland, Ukraine and other countries. (e-Estonia, 2019)

Below is a figure illustrating the as-is situation of data deliveries and frequencies in SE. The number of red balls in the figure illustrate the number of data deliveries that are done by x-road.

**Figure 1. Data deliveries and the number of x-road captures**



The process of managing administrative data involves five departments in Statistics Estonia and goes through the GSBPM phases Specify Needs, Design, Build, Collect, Process, Analyse and Metadata Management/Quality Management (UNECE, 2019).

**Figure 2. The administrative data management process in Statistics Estonia**



The most used administrative source in SE is the Tax and Customs Board. They are the owners of several state registers, and SE captures 80 different datasets from them every year. The frequency of data capture varies from once a day to once a year.

The Population Register is a very important source for population and social statistics. We capture data from the population register through X-Road services, and all the changes are usually captured daily.

For agricultural statistics, the most important source is the Estonian Agricultural Registers and Information Board. At the moment, we capture data every month and have over twenty different datasets in use.

**2. Specify Needs (1) and Design (2) Phase**

*2.1. Administrative data management as-is and challenges*

According to GSBPM, the Specify Needs phase is triggered when the need for new statistics is identified. In SE, the Statistics Design Department gets the input for new data needs from statistical domain departments (Economic and Environmental Statistics Department, Population and Social Statistics Department, Enterprise and Agricultural Statistics Department). This means that the analysts who specify the need for new data consult with the methodologists in the Statistics Design Department, who help to find the data that are already available in SE or confirm that the data needs to be requested from the owner of the register.

According to SE’s strategy, it is important to reduce the administrative and response burden by getting hold of new data sources. This is why we are constantly looking for new sources – to avoid collecting duplicate data at state level. Our analysts analyse the data collected with questionnaires and seek alternative sources, which involves negotiations with data owners to find out how and what we could capture from the source.

Some new data needs and changes in cooperation agreements and data structures are also triggered by developments in the legal environment. For example, new services were recently added to social services, which meant we had to renew our data delivery agreement and X-Road service.

Data delivery agreements are legal contracts between SE and the data owner; they specify deadlines, datasets and data formats, and also general rights and obligations.

In Estonia, there is the administration system for the state information system (RIHA). According to law, it is necessary to get the approval of Statistics Estonia, Data Protection Inspectorate, Estonian Land Board, State Information Agency and National Archive before establishing a database or changing the composition of the data collected in a database. RIHA should give an overview of the information systems and databases that make up the state’s information system, of the data that are collected and processed, of the information systems and services, including X-Road services, that are provided, and the users.

Unfortunately, RIHA does not currently meet all the expectations. After the implementation of the new RIHA version at the beginning of 2018, it serves more as a catalogue of registers and databases and does not give an overview of the data, because descriptions are obsolete or missing. This means that we are not able to receive up-to-date information about changes in the information system’s content.

After getting input from statistical domain departments, the Design phase is carried out by the Statistics Design Department. The Design phase for administrative data includes defining the variables that need to be captured from the data source, compiling information for the data request or contract and preparing the data requests and contracts. In this phase, most of the communication and negotiation with data owners takes place. The administrative data manager’s role in this phase is similar to that of an intermediary or a “translator” – it is important to define the data needs as clearly as possible.

Preparing data requests and cooperation agreements can be challenging because database systems are often complex and SE does not always know their details. Inconsistency in interpreting data request may lead to inappropriate and/or incomplete datasets. For instance, SE has experience with incomplete datasets from the population register. Two supposedly equivalent data requests for birth records gave different results – both had records that were not captured with the other request. Also, when asking for a mother and father for each person, we did not get data on the adoptions of same-sex couples. We became aware of the fact that data on same-sex parents exists in the population register somewhat randomly, while discussing methodology with the representative from the population register. (Statistics Estonia, 2016)

Renewing cooperation agreements can also be time-consuming due to many different reasons. Firstly, specialists of SE constantly find new datasets to research from different registries, which makes the negotiation process very long and assigning the agreement may be postponed due to constantly developing data needs. Secondly, due to developments of information systems and X-Road services, the delay may come from the register’s side.

These examples show that building close partnerships with data owners is the key element in Specify Needs and Design phases, because the quality and relevance of the data received depend on clearly defined needs and unambiguous interpretation by the data owners.

Administrative data management in the Design phase includes describing metadata for administrative data centrally, in cooperation with the owners of registers and statistical domain departments.

Statistics Estonia has a metadata driven statistical information system (SIS). Metadata management system is the main component of SIS. For managing meta information, metadata management system called iMeta has been developed. The main role of the meta information system is to drive the whole statistical process based on the described metadata, including statistical activities, variables, classifications and code lists.

The Neuchâtel terminology model (Neuchâtel Group, 2004) has been used for describing the variables in our metadata management system. In this model, the variables are described in three levels – conceptual variable, statistical characteristic (object variable) and contextual variable. Statistical unit type is an entity for which information is sought and for which statistics are ultimately compiled. Statistical characteristic is a characteristic of a statistical unit type. Conceptual variable (concept) provides a general description of the meaning of the statistical characteristic without explicit reference to any particular statistical unit type. Contextual variable describes the variable in the context of a statistical activity. Contextual variables can be defined as register variables or cube variables.

All administrative microdata captured by SE are described in the iMeta system. This enables the data to be reused in all processes of statistical production. For example, if one data source is used in five studies, the metadata will be described only once. The described metadata will be available directly for the users and for the systems in the live production environment. Describing the metadata centrally enables to harmonise and standardise it.

As regards classifications, Statistics Estonia has to approve the classifications used in the databases that are part of the state information system, and state registers should use the agreed classifications. Harmonisation of classifications and terminology used in administrative registers is one of the key activities in using administrative data in the statistical production process. Standardised metadata specifications will increase the efficiency of data processing.

Managing classifications at the state level is challenging at the moment, because the information about classifications in RIHA can be viewed, but corrections, additions and changes are not visible. Administrative registers currently use their established classifications and definitions, which are often different from the internationally harmonised classifications and definitions. Furthermore, some of the classifications used in administrative registers have not been specified or registered in RIHA.

Describing metadata for administrative data can be challenging, because metadata are not transmitted with data and detailed descriptions are not included in RIHA. Often the definitions of concepts in registries differ from the needs of the statistical production process. Furthermore, most of the registries are technologically managed by large centralised state-level IT-centres, which means that often it is challenging to find the contact person who would know the content of the data in detail.

The wide use of administrative data in SE has produced a lot of information related to data sources. For example, information about cooperation agreements, data requests, data delivery deadlines, data structures, formats, additional information about data, communication with data owners, process phases, etc.

The deadlines for data transmission in SE are currently managed and visualised in the web application JIRA. JIRA enables to monitor the process of data deliveries, data loading, processing, etc. There are different tasks for every data delivery, and every task and subtask can be assigned to a different person. Whenever problems or obstacles arise in some process phase, the questions and answers are inserted in JIRA as comments. This enables to get an overview of the workflow related to the specific dataset.

In SE, the web platform Confluence is used to manage internal information to make it accessible to other colleagues. Every team has its own space or page in Confluence and different overviews and guidelines can be stored and shared that way. We also store information about data sources in Confluence. An overview of administrative data in Confluence consists of a list of institutions and their registries, which Statistics Estonia has a contract with or from whom data are obtained through data requests. In the case of a contract, the date of signing and completion of the contract is attached to it. Each data source also has information about contact persons of the institution to whom it is possible to turn to with data transfer issues. Managing information-related data in SE is challenging, because at the moment the information is stored in different applications and locations.

*2.2 Administrative data management to-be*

Describing high quality metadata for administrative data is the key element of using and comparing data. To achieve the goal of the metadata objects lifecycle and versioning, we are planning to implement the Data Documentation Initiative (DDI). DDI is a free standard for documenting and managing different stages in the research data lifecycle, such as conceptualisation, collection, processing, distribution, discovery and archiving. Documenting data with DDI facilitates understanding, interpretation and use – by people, software systems and computer networks. (DDI Alliance, 2019)

At the moment, we have started to implement DDI by using Colectica software. First, we are planning to migrate statistical classifications from our current iMeta system to Colectica, and then create a webpage view of the classifications for the users. After statistical classifications, other metadata objects will be migrated to Colectica step by step.

One of our future developments for administrative data management is to implement questionnaire-based data capture. We would like to publish on our webpage the datasets that need to be delivered by data owners. The process would be similar to that of questionnaires, where the obligation to fill out the questionnaires is presented on our webpage. This would lower the administrative burden of managing contracts and annexes of contracts. The data structures, formats and deadlines with variable-level descriptions would be published on SE’s webpage from our metadata management system.

In order to enhance the cooperation with the data owners, SE is considering appointing statistical contact persons for each data source. The experience of Statistics Finland shows that statistics contact persons are needed to monitor developments in the field concerned and to work towards maintaining and improving the statistical applicability of register data (Statistics Finland, 2004).

SE has created a vision document of the new information system for managing information related to administrative data. The new information system Administrative Data Gate would enable the data owners to submit different data files through a single entry point. Automated data delivery calendar and reminders would also be part of the Administrative Data Gate’s functionalities.

In the near future, SE will have a central role in the state-level data governance. Part of the state-level data governance are the planned RIHA developments, which means that RIHA will become a state-level data catalogue, where all the data will be enriched with metadata at variable level. These developments will resolve our challenges with poor metadata input. They will also enable SE to get a detailed overview of the data that are available for use at state level.

**3. Build (3) and Collect (4) Phase**

*3.1. Administrative data management as-is and challenges*

Build and Collect phases for administrative data are the responsibility of the Data Service Department. In these phases, pre-processing the data and making them available to the NSI’s in-house applications is the role of administrative data managers. It is ensured through these procedures that there are no duplicate data and that the data are ready for statistical analysis.

Administrative data are captured through different channels:

1) encrypted .csv or .xls(x) files by e-mail, FTP or cloud services;

2) X-Road services that are divided into:

* pull services – the data owner has developed an X-Road service the content of which is suitable for SE. The data are pulled to SE through the X-Road service.
* push services to xGate – the data are pushed to SE through our xGate service. This is the preferred channel for data capture, because SE validates the received data against XSD, and the data delivery process is controlled by SE.

When administrative data have been captured through different channels, the loading processes begin. The first step is loading the data to the Initial Observation Registry (IOR). When the data are sent by .csv or .xls(x) files, the data will be loaded to Oracle database as they arrive. Loading and processing the data that has been sent with files is time-consuming for us, because there are constant problems with agreed data structures and wrong data formats.

When data are captured by X-Road pull services, the XML file is parsed to the IOR by Oracle tools. When data are captured by xGate, the file is parsed and validated against the XSD file generated in the iMeta system. After loading the data to IOR, it is possible to give the first feedback about the received data. The captured data are unloadable if the formats are incorrect or there are missing variables.

The next step is Data Staging Area (DSA), where data structure checks and conversions to correct formats take place. These checks and conversions are done according to the metadata descriptions in iMeta. It is also possible to develop more contextual checks, but for this, the input for the rules is needed from statistical domain departments. After DSA, it is possible to automatically generate a quality report about the delivered dataset.

The last step is to make the data available for users, which means that the data are loaded to Final Observation Registry (FOR) and are pseudonymised if the data include personal data. The process of pseudonymisation involves removing personal identification numbers, names and contacts from the data. PIN-numbers are replaced with unique identifiers that allow the data to be joined. The unique in-house identifiers are not derived from PIN-numbers, which means that it is not possible to convert the unique identifiers mathematically to PIN-numbers.

**Figure 3. Administrative data flow in Statistics Estonia**



The data are stored and versioned in Oracle databases, which are available for use to statistical domain departments through SAS or R.

*3.2 Administrative data management to-be*

The to-be plans for administrative data management in Build and Collect phases include implementing more automated controls in DSA. It is also SE’s strategy to move as much data deliveries as possible to X-Road push services via xGate.

SE has created a vision document to develop new information system Administrative Data Gate. It will help automate the administrative data management in Design, Build and Collect process phases.

The main functionalities of the Administrative Data Gate are:

* Monitoring data deliveries and sending automated feedback and reminders to data holders.
* Reading metadata from SE’s metadata management system and checking delivered data against the agreed structures and content.
* Functionality to convert data to formats or structures needed by statistical domain departments.
* Administrative Data Gate will allow to log and monitor every procedure that is done with the specific dataset.
* Dashboard with main operations visible for users.

**4. Process (5) and Analyse (6) Phase**

*4.1. Administrative data management as-is and challenges*

Process and Analyse phases involve data quality checking, which is done in SE by statistical domain departments. In these process phases, the analysts who check the delivered administrative data contextually are in the administrative data manger’s role.

Checking data quality beyond the obvious (e.g. missing values, unique identifiers) requires creativity. The analysts in statistics domain departments have to compare data of different time periods and from different sources to track quality problems. Comparisons are done with SAS or R tools, and methodologies for checking quality depend on the specific source or dataset.

As Estonia is planning to conduct its first register-based census, the quality of the registers in scope have been analysed most thoroughly. The index-based methodology of the census gives a different view of the known datasets and allows to compare multiple sources. However, getting a grasp of systematically missing data when no comparison is available remains challenging.

Statistics Estonia has issued requirements for registry holders, which the latter have to meet before the start of the register-based census:

• at least 97% coverage of the population;

• information on required census characteristics must be regularly updated and this process must be documented;

• at least 95% of entries must have an identifier in the standard format;

• at least 95% coverage of relevant census characteristics;

• the rate of material or technical errors in the values of relevant census characteristics may not exceed 1%. (Statistics Estonia, 2016)

The 25 registers that are in the scope of the register-based census have been evaluated towards these requirements at the first pilot census and will be evaluated again during the second pilot census.

*4.2 Administrative data management to-be*

The to-be plans for administrative data management in Process and Analyse phases include implementing more automated contextual data checks. Some of the comparisons and data quality checks could be automated and are planned to be implemented in the Administrative Data Gate information system.

There are also some conditions that have to be met to be able to use administrative data in the statistical production process:

1. Estonian addresses in all state-level information system (registers) are extracted from the address data system of the Land Board – this ensures standardised address data;

2. All persons in databases are identified on the basis of personal identification code (for natural persons) or registry code of the commercial register or non-profit associations and foundations register (for legal persons);

3. The data for register-based census and other surveys are captured according to the agreed data content and structures, which allow the performance of statistical activities complying with the quality criteria of official statistics.

State-level data governance will also ensure that the conditions will be met in all state registries.

**5. References**

Eurostat (2018), Good practices in accessing, using and contributing to the management of administrative data <https://ec.europa.eu/eurostat/cros/system/files/admin-wp1.2_good_practices_final.pdf>

DDI Alliance (2019), DDI Alliance’s webpage <https://www.ddialliance.org/>

UNECE (2019), Generic Statistical Business Process Model v5.1 [https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1](https://statswiki.unece.org/display/GSBPM/GSBPM%2Bv5.1)

Statistics Estonia (2016), Improvement of the use of administrative sources (ESS.VIP ADMIN WP6 pilot studies and applications), Methodological report <https://ec.europa.eu/eurostat/cros/system/files/admin_wp6_2016_ee.pdf>

Statistics Finland (2004), Use of Register and Administrative Data Sources for Statistical Purposes. Best Practices of Statistics Finland. [http://www.oecd.org/sdd/36237299.pdf p.21](http://www.oecd.org/sdd/36237299.pdf%20p.21)

E-estonia (2019), E-estonia’s webpage

<https://e-estonia.com/solutions/interoperability-services/x-road/>

Neuchâtel Group (2004), Neuchâtel Terminology Model for classifications (version 2.1) and variables (version 1.0)