**Data Architecture for Statistics Production – Logical Data Repositories**

Tuukka Saranpää, Statistics Finland, tuukka.saranpaa@stat.fi

**Abstract**

*Traditionally, data repositories at Statistics Finland have been designed for the need of one or a few specific statistical areas. Additionally, development projects on data repositories have typically assumed a systems-oriented approach, focusing on the technological solutions required. Over time, the lack of top-down coordination has resulted in overlap, inconsistency, poor opacity and impaired quality between different data repositories. These issues, along with new national guidelines for Finnish public administration, prompted Statistics Finland to organize a preliminary study to outline the fundamental requirements for a common statistical data architecture as well as a roadmap for its development. In January 2017, following the findings of the preliminary study, Statistics Finland launched a development program on data architecture for statistical production to coordinate the various projects outlined in the roadmap. In order to ensure consistency and interoperability between different development projects, a top-level to-be model for logical data repositories for statistics production was required. In developing the model, Statistics Finland assumed a holistic top-down view on statistical production, shifting the focus from statistics to data. The model envelops the entire statistics production process (Generic Statistical Business Process Model phases 4-7), from data acquisition to dissemination and determines the names and general constitution of all relevant top-level data repositories. Some key observations of the new model are the emphasized role of metadata through the process and a centralized top-level repository for geospatial data serving all areas of statistics production. The newly-developed model provides a top-level framework for any development effort on a specific area. However, to truly capture the benefits of a common data architecture, significant bottom-up development work is required to determine the exact content and physical infrastructure of each repository. Furthermore, the top-level model may be improved based on observations and experiences from lower-level development.*

**Keywords:** Data architecture, Statistics Production, Logical data repositories

**1. Introduction**

One of the big questions for any statistical organization is how to organize the vast amounts of data acquired, utilized and created in statistics production. Data organization decisions and solutions may have a significant impact on the production process, including aspects such as accessibility, usability, opacity, consistency and quality. Traditionally, statistical organizations have been focused on statistics and consequently, the defining logic for organizing data has been from the viewpoint of statistics. This paper, however, proposes that in order to avoid overlap and improve quality, statistical organizations need to focus specifically on the qualities of the data when designing their data architecture.

This paper presents the data architecture development work at Statistics Finland, focusing on a newly-developed model for logical data repositories for statistics production. The paper is organized as follows. In chapter 2, we present the background for data architecture development at Statistics Finland. Chapter 3 presents the top-level to-be model for logical data repositories, while chapter 4 discusses its potential implications for quality in statistics production. Chapter 5 concludes.

**2. Background**

Statistics Finland has a long history of developing data repositories for statistics production. Traditionally, these repositories have been designed for a specific and restricted purpose, i.e. one or a few statistical programs. Development of data repositories has historically also been somewhat systems-oriented, focusing primarily on the software and technology solutions. While definitely improving and unifying the production process in the target statistical programs, traditional data repositories have not been able to overcome larger-scale issues, such as overlap, poor opacity and inconsistency of data between different data repositories and/or statistical programs.

Shortcomings of the traditional data repositories, along with new national guidelines for Finnish public administration, prompted Statistics Finland to launch a preliminary study for the development of data architecture in late 2015. In order to approach such a large-scale and comprehensive topic, the preliminary study was tasked with (1) identifying the issues and shortcomings of the current state, (2) outlining the requirements for a common statistical data architecture, and (3) defining a roadmap for its development. The roadmap identified a number of development projects focusing on specific areas of statistics production for 2016 to 2018. Importantly, however, the preliminary study also underlined the need for producing high-level top-down data architecture definitions to guide and assist all lower-level specific development work.

In January 2017, following the findings of the preliminary study, Statistics Finland launched a two-year development program on data architecture for statistical production. The primary purpose of the program is to coordinate and facilitate interaction between the various projects outlined in the development roadmap. In addition, the program has launched a number of sub-projects to produce the top-level data architecture definitions and guidelines required for overall consistency and interoperability.

The first sub-project of the program served to create the foundation of data architecture development at Statistics Finland. Following the guidelines of Enterprise Architecture for Finnish public administration, the project produced a number of definitions and depictions for top-level data architecture. The project work started at the conceptual level by forming an understanding of the concepts and terms used within statistics production. An important part of this effort was adopting the Generic Statistical Information Model (GSIM) as the basis for determining a local conceptual model for Statistics Finland. At the conceptual level it was also established that the data architecture design should, by nature, be data-centric instead of statistics-centric.

From the conceptual level the project moved on to study the top-level logical data structures required for statistics production. Here, the data-centric approach allowed for an open-minded approach to the actual data and data entities resulting in a top-level to-be model for logical data repositories for statistics production. The model and its implications are discussed in the following chapters.

**3. Logical data repositories**

The top-level to-be model for logical data repositories for statistics production is presented below in Figure 1. The model depicts all logical data repositories required at the different phases of the statistical production process within the context of Statistics Finland. However, the model aims to be generalizable – for example, in the future it could be expanded to cover all official statistics production in Finland.

The titles of the logical repositories – translated to English for the purposes of this presentation – are descriptive of the content of the repository and thus do not represent the underlying data systems or IT solutions. It should also be noted that the logical data repositories are not designed to directly determine the systems architecture – one logical data repository may require several different systems solutions or one systems solution may serve several logical data repositories at different phases of the process.

The first key observation of the model is that it envelops the entire statistics production process from data acquisition to dissemination (based on the Generic Statistical Business Process Model, GSBPM, officially adopted for use at Statistics Finland in 2015). Traditionally, the data repositories at Statistics Finland have focused on data processing and analysis, resulting in a gap between data acquisition on the one hand, and dissemination on the other hand. The to-be model is the first concrete representation of (logical) data repositories through the complete data value chain. However, while the model does recognize the role of data repositories at the collection and dissemination phases, the internal structure of these repositories is currently undefined. Projects for further defining the contents of the logical data repositories are currently underway.

**Figure 1. Logical data repositories for statistics production**



Another key observation of the model is the emphasized role of metadata. The model aims to communicate the significance of metadata at all phases of the statistical production process from data acquisition to dissemination. Furthermore, the model facilitates a lifecycle perspective on metadata, where (1) the initial metadata is acquired (and potentially converted) at the data collection phase, (2) metadata is created, refined and supplemented at all phases of the production process, and (3) metadata is finally refined and disseminated to end-users. The emphasized role of metadata in the data architecture also supports the design and improvement of the business architecture allowing for new metadata-driven applications, services and processes.

The third key observation of the model is its data-centric approach. In the current top-level model, this is most apparent in the logical data repositories for the processing phase of statistics production. Here we have defined six top-level logical data repositories divided by their content and not by their intended use in specific statistical programs. This approach promotes the joint use of data through accessibility and opacity between statistical programs while discouraging unnecessary copying by storing each data point only once. The advantages are discussed in more detail in the following chapter.

The top-level logical data repositories are large and abstract entities and thus offer very little concrete information on how the data is organized at a lower-level. Therefore, lower-level development work is essential in defining the data architecture for each top-level data repository. For example, another development project at Statistics Finland has further divided the top-level “Population and Social Data Repository” into nine sub-repositories grouped by their data content, such as Education Data or Health Data. Importantly, the top-level model provides all lower-level development efforts a map or framework for identifying their position in the overall data architecture as well as their connections to other repositories.

In addition to providing a top-level framework, the to-be model can also be utilized to recognize important top-down areas of development. For example, identifying geospatial data as a top-level logical data repository has allowed Statistics Finland to organize a specific development project for defining the data structures and solutions required for providing geospatial services for statistics production.

For effectively developing the NSI’s data architecture it is important to apply and integrate both top-down and bottom-up approaches. While the top-level logical data repository model provides Statistics Finland with an overall understanding of the required data architecture, a thorough study of data utilized in the production of over 70 statistics was required to determine the exact sub-structure of a single top-level data repository, namely the Population and Social Data Repository. It should also be noted that any data structure does not exist in a vacuum – instead it should be closely linked to the business architecture, that is, actors, services and processes of the organisation.

**4. Implications of data architecture for quality in statistics production**

The data-centric approach to designing the logical data repositories presents numerous potential implications for quality in statistics production. However, as the development work at Statistics Finland is still at a fairly early phase, the impact of the approach is, for the most part, based on assumptions. Additionally, the potential benefits will not be actualized until IT systems and solutions corresponding to the logical guidelines have been developed and taken into use.

According to the preliminary study on data architecture, many of the issues with the current state stem from producing multiple copies of the same data, poor visibility between data systems and a statistics-based perspective on the use and ownership of data. The new data-centric approach, on the other hand, allows the data architecture design to focus on data, the qualities of data, the flow of data between repositories and the (joint) use and reuse of data.

Providing a logical top-level overview of data repositories, along with assigning each data repository a logical and descriptive title, improves opacity within statistics production. By mapping the physical data systems to the logical data repositories, it is simple to deduce the location and nature of any data. Ample and high-quality metadata, on the other hand, assists in interpreting and assessing the data. A commonly shared conceptual model ensures that data is structured similarly in each data system.

A significant aspect of the data-centric approach is the joint use of data. In the to-be model, data is not collected, processed and analysed from the viewpoint of a specific statistical program. Instead, each process phase should be assessed holistically from the viewpoint of overall statistics production. Data flows are not determined by statistical programs (“statistics pipeline”) – instead, data is semantically directed into the corresponding data repositories and processed for general utilization of statistics production. Data governance and administration allow statistical programs to access the data they require for production. Not only does the data-centric approach discourage unnecessary copying of data, it also improves consistency through the use of shared data – the data is checked and processed from more than one viewpoint, as the organization is aware of how and where the data will be used. A holistic viewpoint on shared data also allows the organization to identify potential overlap or recognize common needs for data collection.

Finally, assuming a data-centric approach allows the organization to see statistics production as merely one – albeit probably the most important – use for data. The data repositories may be similarly utilized for other information products, such as preparing micro data sets for research purposes. When the data is structured, shared and accessible with ample metadata available, the organization may also identify completely novel uses for the data, including new data products to serve the end-users.

**5. Conclusion**

Based on the experiences from early development work, we suggest that assuming a data-centric approach to defining a top-level data architecture may serve as a solid foundation for improved consistency, opacity and quality in statistics production. However, in order to achieve the desired impact the top-down approach should be combined with bottom-up development interactively – the principles and definitions from the top guiding the lower-level development and findings from the lower-level work serving as feedback for improving and defining the top-level models. Furthermore, concrete results will not be achieved until functioning physical solutions and IT systems are developed consistently following the definitions and principles of logical data architecture.

Statistics Finland is currently carrying out and planning several development projects for producing a more defined data architecture for each target area. Until the end of 2018, the data architecture development program will coordinate and facilitate interaction between the various projects. However, we are currently also designing practices and organizational structures for making data architecture an inherent part of Statistics Finland. Another aspect of development in the next years will be moving towards shared data architecture practices between organizations of Finnish public administration.