

25<sup>th</sup> Meeting of the Wiesbaden Group on Business Registers  
- International Roundtable on Business Survey Frames

Tokyo, 8 – 11 November 2016

*Irene Salemink*  
*Statistics Netherlands*  
*Session No. 5*

*Technology*

***A Unit-base derived from the Statistical Business Register as springboard to a Data-lake***

**Abstract**

The strategic agenda of Statistics Netherlands (SN) comprises 10 goals of which one states “Towards a state of the art data and information architecture”, that also includes “To make data more and better accessible for compilers of statistics by implementing a Data-lake”. Where the Data-lake is outlined as: a concept that foresees in a decoupling of input, statistics production and output activities, satisfies the demand for flexibility and coherence and also enables both statistics compilers and users to acquire their information as self-reliant as possible without the necessity of being a methodology- and/or IT-expert. At SN this concept is not confined to the technical approach how to store and retrieve data, it is in fact “trade mark” for all aspects of future focused data management.

This paper describes the approach and the steps SN is undertaking in the process of realizing a Data-lake by defining; the concept from the user perspective, the expectations of users, the foreseen use, the proposed new business architecture and the foreseen capabilities of the Data-lake. Also the approach of user stories based on real examples that are used to elaborate and examine the capabilities will be explained. An important role in this concept is reserved for the Unit-base that is derived from the Statistical Business Register (SBR).

With the introduction of the Data-lake concept a first step is taken towards a new Business Architecture in which data has a central role and where the processes are positioned around the data; a data driven architecture. This in contrast to the current process driven architecture where the processes are centralized and data is stored in “steady states”. A data driven architecture demands a different way of data handling; the data is made available centrally and virtually and the processes will become metadata driven. The foundation on which the Data-lake is build is the so called Data-source layer of which the Unit-base is one of the key building blocks.

This Unit-base has been recently developed (June 2016) and is currently in place and operational. The Unit-base is derived from the SBR and contains information on populations of statistical units as well as administrative units and their data. In the way how the Unit-base is created, it’s positioning and use in the statistics production process and the retrieval of data from it by users (by making use of ready-made building blocks and webservice technology), the Unit-base is a simplified example of a functioning Data-lake concept. Thereby empowering the role and use of the SBR as backbone for statistics production and providing a solid base for further investigation and development of a full blown Data-lake.