26th Meeting of the Wiesbaden Group on Business Registers
- Neuchâtel, 24 – 27 September 2018

Irene Salemink
Statistics Netherlands

Integrated Statistical Register Systems

How the Data lake approach can strengthen the SBR role and vice versa

Abstract

Statistics Netherlands' mission is to publish reliable and coherent information that adapts to the need of society. The demand for information on the one hand asks for fast and flexible information with high quality. On the other hand there rises a social necessity to join one's own data with data of SN. To meet both requirements SN wants to be able to access, share, combine and re-use data, without endangering privacy sensitive data. In order to meet this challenge SN is developing a "Data lake" solution.

The rationale behind the Data lake concept is to "connect" statistical processes to the Data lake in order to access the potential of reusable data. This implies that the Data lake should contain all usable data, at least from a consumers point of view. In real the data are not stored physically in one place (e.g. a database). The SN Data lake is a so called logical data warehouse. In order to create the look and feel of "data at your fingertips" SN is applying data virtualization, metadata-modelling and semantic technologies.

The concept of sharing data applies for both internal (statistical) data, on premises, as well as data stored externally at the source owners' location. The Data lake contains functionalities to enable faster and easier data handling, like for example searching, finding and understanding data, as well as for deriving, combining and transformation of data. Also logging, monitoring, authentication and security functionalities are foreseen.

A special place in the Data lake is kept for the "source layer", the holy grail where virtually all data sources are "accessible". Next to all kinds of statistical- and register data also the Statistical Business Register can be approached as a data source. The Data lake approach facilitates to access the SBR-data as well as to join and combine various (register) sources with it, in a virtual manner, thus without copying, moving, duplicating or transforming data physically. This concept was tested in various Proof of Concepts of which the implementation of the use case "Family Businesses" was the pièce the resistance thus far. The basic idea to detect a FB in the SBR is to link additional administrative data to the existing administrative and statistical units stored in the SBR. In this way a 'satellite' is created to characterize a FB. For this satellite no less than 8 sources need to be coupled and disseminated. The 16 views comprising the methodology behind FB were created in less than 40 minutes.

Keywords: Data lake, meta data, data virtualization, Statistical Business Register, Family Businesses