Quality Guidelines for the Acquisition and Usage of Big Data with additional Insights on Web Data

Alexander Kowarik, Magdalena Six
Statistik Austria, Vienna, Austria.

Abstract

The increasing knowledge and experience within the European Statistical System (ESS) in the acquisition, processing and use of new data sources provides now a clearer picture on quality demands. These guidelines use the quality based experiences in the ESSnet Big Data II to formulate guidelines for NSIs who already use and/or plan to use new data sources for the production of Official Statistics. Looking at the production process of statistics, the usage of new data sources mostly affects quality aspects of processes related to the input and the throughput phase. Taking this into account the guidelines concentrate on the input and the throughput phase of the statistical production process.

With new data sources, the access to as well as the processing of input data makes it necessary to consider new and very source- and data-specific sub-processes. The variety of sub-processes is much broader compared to the use of traditional data sources. What is relevant for one data class and one data access might be of no interest for others.

The Web Intelligence Network (WIN) builds on the work of the previous ESSNets Big Data I and Big Data II and adapts and expands the focus on the more specialized usage of web data.

In this paper we describe the outline of the formulated Quality Guidelines as well as the challenges to structure the wild field of new (sub-)processes and aspects when new data sources are used in the production process of Official Statistics. We give examples of specific quality guidelines and further, we risk an outlook how quality guidelines will develop when the usage of new data sources has become a normal part of the production process of Official Statistics.