Quality Aspects of Web Data based on the Experiences of the ESSnet Trusted Smart Statistics – Web Intelligence Network

Author: Magdalena Six

Coauthors: Alexander Kowarik

Submission ID: 555

Format: CPS Paper

Reference Number: 555

Presentation File

abstracts/ottawa-2023_37bbb5d30544becb6e238baaf2c7ec3f.pdf

Brief Description

Web data for the production of official statistics

Abstract

The ESSnet “Trusted Smart Statistics – Web Intelligence Network (WIN)” is a project within the European Statistical System (ESS), which engages 17 organizations from 14 European countries. It aims to develop a web intelligence system at the ESS level, providing a greater chance to generate the right conditions for the integration of web data into official statistics.
One work package (WP2) of this ESSnet includes already well-established use cases such as online job advertisements (OJA) and online-based enterprise characteristics (OBEC), with the ambition to be moved into the statistical production stage soon. Another work package (WP3) focusses on new types of web data sources, such as web data about the real estate market, construction activities, online prices or hotel prices. For these use cases the aim of the ESSnet is to produce experimental statistics.
Building on the experiences made in the different use cases, a work package of its own (WP4) aims to consolidate knowledge gained in the WIN in the area of methodology, architecture and quality when collecting, processing and analysing web data. Based on the more generic work of previous ESSnets (Big Data I & II), the members of WP4 already collected and published “Minimal guidelines and recommendations for implementation” w.r.t. quality, methodology and architecture. The included quality guidelines are structured along two phases of the statistical production process, the input phase and the throughput phase. The throughput phase refers to two different processes and is split into two parts. The first part of the throughput phase is dedicated to deriving- so-called - statistical data from web data. In this part we pay particular attention to quality aspects such as linking, coverage, measurement errors and model/processing errors and comparability over time. The second part deals with usage of the derived statistical data to produce statistical output.

During the duration of the ESSnet, the quality guidelines will be systematically extended with input from the use cases in WP2 and WP3.

In the paper, we will discuss the structure of the already published quality guidelines, introduce some examples and focus especially on the new developments from the ongoing ESSnet.