UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Collection and Sources
22-24 May 2024, Geneva, Switzerland

19 April 2024

Tapping into web data for European statistics – challenges and experiences of the ESSnet Web Intelligence Network

Klaudia Peszat and Dominika Nowak (Statistics Poland, Poland)

k.peszat@stat.gov.pl

Abstract

The experiences of many National Statistical Offices have provided evidence of the relevant role of web data in producing new and augmenting existing statistics. However, the integration of web data with official statistics is a demanding process and the quality of the output very much depends on the quality of the source. Thus, transforming the information available on the web into statistical data requires significant methodological work. Within the ESSnet Web Intelligence Network project the partnership of 17 organizations explore the potential of web-scraped data for the production of European statistics in several domains, such as online job advertisements, online-based enterprise characteristics, real estate market data, construction activities, online prices for household appliances, tourism, and business registers enhancement. The current experience demonstrates the challenges related to data acquisition and processing steps, which cover landscaping of web data sources relevant for the topic of interest, the stability of web sources, technical aspects of web scraping, dealing with deduplication, annotation of data sets, ensuring the quality of classification models, etc. Additional issue is the automation of data collection processes in the Web Intelligence Hub – the platform for central web scraping, developed by Eurostat. This paper discusses the selected issues related to the use of web data sources in official statistics in the context of developing a universal tool enabling the acquisition, processing and analysis of web data at the European level, i.e. WIH.