# Exploration and experience with new web data sources. A Case Study for innovative tourism statistics

**Galya Stateva[1], Marek Cierpiał-Wolan[2]**
Bulgarian National Statistical Institute, Bulgaria, [2] Statistics Poland, Poland.

***Abstract***

*The aim of the first part of presentation is to tap into the potential of new web data sources, which will have the potential to be integrated in the Web Intelligence Hub, developed by Eurostat. Parallel to the exploration of the data sources, we aspire to produce experimental statistics, using these new web data sources, given that they meet the quality criteria.*

*The presentation will delve deeper into Work Package 3, part of the European Statistical System Collaborative Network (ESSnet) Web Intelligence Network (WIN) project, dedicated to the exploration of non-traditional data sources for official statistical production.*

*Work package 3's activities are divided into six use cases, each having distinct characteristics and specific goals:*

*• Use Case 1 aims to explore new data sources and monitor the real estate market.*

*• Use Case 2 aims to derive early estimates of construction activities, pertaining to both already built and planned buildings, based on real estate web portals.*

*• Use Case 3 aims to collect data about online prices of household appliances and audio visual, photographic and information processing equipment by web scraping of online shops and at a later stage compare the data with scanner data for the shop's sales.*

*• Use Case 4 aims to develop new indices for tourism statistics, using the data from booking portals, air traffic portals, travel agencies portals and portals related to quality of life.*

*• Use Case 5 is concentrated on mass web scraping, primarily for the enhancement of the quality of the business register via linking URLs of enterprises and predicting main economic activity codes (NACE)*

*• Use Case 6 aims to explore the use of publicly available traffic camera data in order to produce new indicators. In this use case a peculiar data source is used – pictures from traffic cameras and induction loops.*

*Use cases 1-4 share similar characteristics in terms of data sources and expected experimental indicators and adhere to pre-defined process steps in compliance with Big data life cycle, which include "New data sources exploration", "Programming, production of software", "Data acquisition and recording", "Data processing", "Modelling and interpretation" and "Dissemination of the experimental statistics and results". Use cases 5 and 6 take a slightly different approach due to their extraordinary data sources and do not adhere to the aforementioned process steps.*

*During the first project's year, the Work package 3 achieved meaningful results, such as a Checklist used as a tool for assessment and justification of web data sources, defined a set of mandatory and optional variables to be extracted from the data sources, sets of minimal indicators, based on the mandatory variables, successfully set up and tested their working environment and software solutions for the upcoming data collection, literature review focused on URL finding methodology and tools and the use of business websites to predict economic activity of enterprises, preparation of training and tests sets and accompanying methodology for URL finding, preparation of the upcoming NACE prediction and classification, exploration of the available assessment of the model results, implementation of Machine-learning pipeline for publicly accessible traffic camera data.*

*We are also scheduled to begin testing of Eurostat's Web Intelligence Hub for specific use cases from our Work package, which volunteered in the endeavor.*

*While we have successfully implemented our initial planned activities for the first project year we continue our work, constantly monitoring the available resources, arising issues and quality of the data, which is to be collected and processed during the second project year.*

*The different use cases have already encountered potential and expected issues like the possible changes in the source of web data structure and web site changes, checks for legal and copyright constraints, non-standard variables, mechanisms blocking extraction of data (e.g. javascript, captchas, etc), viability of training and test sets for both URL finding and NACE prediction, difficulties when comparing results with other partners, since NACE code classification is knowledge-intensive and language-specific*

*sources have to be used, regular update of the data source. Due to the peculiar data sources for some use cases we have also encountered unsolvable issues like weather variation (e.g. snow,rain, darkness). Some of the issues have been solved, while others still remain.*

*A Case Study for innovative tourism statistics aims to show the achievements of two projects: ESSnet Big Data II and ESSnet WIN concerning the use of unstructured data sources in the field of tourism.*

*The work in the Big Data II project started with an inventory of data sources related to tourism statistics, which can be used for research of tourist accommodation establishments as well as for estimating tourist traffic and related expenditures. The VisNet tool was developed to visualise the links between the identified sources.*

*The gathering of data from digital sources required the preparation of a scalable solution for data retrieval using web scraping techniques. The developed author's method allowed for continuous and non-invasive extraction of data from selected accommodation booking portals.*

*The process of integrating statistical databases with data derived from web scraping required the development of a fully automated innovative tool, which unified the structure of identification data and assigned them geographical coordinates. The preparation of appropriate structures allowed the implementation of methods of combining data from different sources.*

*The project also developed a methodology for estimating the volume of tourist traffic and tourist expenditures using spatial-temporal disaggregation methods or the method of flash estimates of accommodation establishments.*

*As a result of the work carried out, a prototype of the Tourism Integration and Monitoring System (TIMS) was prepared, together with dedicated micro services, which will support statistical production in the area of tourism statistics and assist in monitoring changes in the tourism sector.*

*The continuation of the work initiated in ESSnet Big Data II is the ESSnet WIN project, in which new methods for assessing the quality of external data sources have been introduced and web scraping has been expanded to other types of portals related to tourism. The main objective of the project is to develop new indicators, which will be an integral part of the developed prototype.*

---