FICHA DE CURSO

Collecting and Using Web Data. Applications to some statistical domains

Fechas: 6, 7, 13 y 14 de junio de 2023

-CURSO IMPARTIDO EN INGLÉS-

Plazo presentación de instancias: del 12 al 25 de mayo de 2023

Lugar de realización: Avenida de Manoteras, 50 - 52, planta 6ª aula 006

Duración: 16 horas lectivas

Horario: De 09:00 a 13:30 horas

Course coordinators:

Subdirección General de Metodología y Diseño de Muestras

Objective:

The course will cover the main steps of web data acquisition and processing illustrated with the use-cases of *Business Registers and Tourism Statistics enhancements with web data*. It will lead participants step by step from assessing the content of the webpage, through creating data acquisition tools and presenting possible uses of the collected data. Methodological and quality-related challenges with the use of web data in official statistics, as well as possible solutions developed by the European Web Intelligence Network, will be discussed. Several techniques of merging datasets from different sources (e.g. web scraping, sample surveys,…..) will be proposed.

The learning objectives will be (i) to share practical aspects of integrating web data with official statistics, (ii) to provide an overview of web data acquisition, processing, and methods of analysis, (iii) to discuss methodological challenges, limitations of the use of web data in official statistics, and (iv) to present possible solutions.

Content:

a. First steps in web scraping in R: websites, text strings, geographical coordinates, record linkage.
b. Use case: accommodation establishments.
c. URL finding, scraping from enterprise websites and processing their web data. Classification technique for NACE prediction.
d. Use case: strategies from the Web Intelligence Network.

Aimed at:

Statistical staff involved in the production of official statistics

Previous knowledge requested:

R basic knowledge.

Trainers:

Johannes Gussenbauer, Statistics Austria. (Member of the Web Intelligence Network ESS-Project)

Sebastian Wojcik, Statistics Poland. (Member of the Web Intelligence Network ESS-Project)