The Use of Web Data for EU Official Statistics. Case Studies on Online Job Advertisements and Online Based Enterprise Characteristics.

Jacek Maślankowski[1]

[1] University of Gdańsk, Poland; Statistics Poland
e-mail: jacek.maslankowski@ug.edu.pl

Abstract

The wide use of Big Data and Smart Statistics services in official statistics is a fact. In this paper we are describing two case studies: OJA – Online Job Advertisements and OBEC – Online Based Enterprise Characteristics. The paper relates to the implementation of Big Data as a source for official statistics within the last 6 years and plans for the next 3 years. First project regarding the OJA and OBEC data collection started in 2016 with the Big Data ESSnet 1st Grant. Then it was continued with the Big Data ESSnet II grant. Currently, under Web Intelligence Network initiative, implementation of the production process is on the way for both OJA and OBEC. We are expecting to move these two use cases into production in 2025.

The analysis of OJA data is well known in research papers. Its form varies from qualitative analysis (Hinrichs, Bundtzen, 2021) to quantitative based on text analysis (Cao et al., 2021). Currently Eurostat is analyzing several million job advertisements (Ascheri et al., 2022), based on the system developed in past by the European Centre for the Development of Vocational Training - Cedefop (Descy et al., 2019). For the OBEC data, more data sources are used, i.e. enterprise websites. The challenge is to define legal and methodological frameworks to be replicated by different NSI's, with respect to the business register data and associated URLs of enterprises.

The OJA project has been initiated to collect of job advertisements published online on the web as a new source of data in the field of labour market statistics. For this project several webscraping procedures were developed and tested, including privacy and legal issues of collecting web data. Another document was "webscraping policy" issued in collaboration with the OBEC project (wpc_deliverable_c1_ess_web-scraping_policy_template_2019_07_15.pdf, europa.eu). The development of OJA indicators and analyses of OJA data is now undergoing and data is presented as experimental statistics. Those data can be accessed via Eurostat OJA DataLab where they are regularly improved and updated. Data quality of the OJA data is necessary to decide what data can be disseminated as a part of official statistics. OJA Datalab contains information on many properties of open job advertisements in the market, and the daily dynamics are much richer than quarterly collected survey data. Thus, the data quality assessment also focused on how OJA data refers to official Job Vacancy Statistics (JVS) (Maslankowski et al., 2022).

Regarding OBEC the main goal is to collect the largest number of URLs to be used in web scraping process. The population of OBEC use case includes enterprises employing 10 or more employees. The population was defined as in ICT in Enterprises survey , i.e. "Definition of the enterprise website, From: Methodological manual for data compilers and users of the ICT survey, A6. Does your enterprise have a website? [Scope: enterprises with access to the internet, i.e. A1 > 0], [Type: single answer (i.e. Tick only one); binary (Yes/No); filter question]" (ESSnet WP-C, 2020). There are several ways to acquire URL list of enterprises. One way is to use official registers which in many cases have this information available publicly. For instance, in Poland there are enterprise registers like KRS or CEIDG, which are available online and it is possible to get the website URL if available for the company. There are also commercial databases with the information on enterprises in specific country. One of them is an ORBIS database by Bureau van Dijk which includes detailed information on enterprises, including URLs for numerous enterprises. But the responsibility for the data quality relies on the third-party vendor which

makes the data difficult to be used in official statistics. It is also important to note that fees apply for data acquisition.

Keywords: big data, machine learning,

References

Ascheri, A., Marconi, G., Meszaros, M., Reis, F. (2022) 'Online Job Advertisements for Labour Market Statistics using R', Romanian Statistical Review, (1), pp. 3–26.
Cao, L, Zhang, J., Ge, X., Chen, J., (2021) 'Occupational profiling driven by online job advertisements: Taking the data analysis and processing engineering technicians as an example', PLoS ONE, 16(6), pp. 1–20. doi: 10.1371/journal.pone.0253308.
Descy, P., Kvetan, V., Wirthmann, A., & Reis, F. (2019). Towards a shared infrastructure for online job advertisement data. Statistical Journal of the IAOS, 35(4), 669–675. https://doi.org/10.3233/SJI-190547
ESSnet WP-C Deliverable C6, Deliverable C6: Reference Methodological Framework for processing online based enterprise characteristics (OBECs) data for Official Statistics V .2, 2020, Available at: https://ec.europa.eu/eurostat/cros/sites/default/files/WPC_Deliverable_C6_Reference_Methodological_Framework_v2.0.pdf
Hinrichs, G. and Bundtzen, H. (2021) 'Communicating a Sales Job to Occupational Changers: A Qualitative Content Analysis of Online Job Advertisements', TEM Journal, 10(2), pp. 853–857. doi: 10.18421/TEM102-45.
Maslankowski et al. (2022), WP2: OJA and OBEC Software, Deliverable 2.1: WP2 1st Interim Progress Report.