

**ESSnet Big Data II**

**Grant Agreement Number: 847375-2018-NL-BIGDATA**

<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata>

[<https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en>](https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en)

**Workpackage WPB**

**Implementation – Online Job Vacancies**

**Home data vs. CEDEFOP data – Case Study for Poland**

**Prepared by:**

**Jacek Maślankowski (PL)**

**Date of present version: 7 September 2020**

Workpackage Leader:

Tomaž Špeh (SURS, SI)

e-mail address: tomaz.speh@gov.si

mobile phone: +38651672116

Contents

[1. General overview of home data 2](#_Toc50367664)

[2. Methodology of data processing and analysis for home data 3](#_Toc50367665)

[3. Main differences between CEDEFOP and home data 4](#_Toc50367666)

[4. Main conclusions and future work 5](#_Toc50367667)

[Figures and tables 5](#_Toc50367668)

[Literature 6](#_Toc50367669)

# General overview of home data

The data for Poland are scrapped every day from largest Job Offers Portals, and occassionally (for a specific publication) from additional data sources. It includes (number of job offers for 2017):

* Commercial data sources (Internet portals)
  + money.pl – 140 thous.
  + gowork.pl – 124 thous.
  + jobs.pl – 46 thous.
  + pracuj.pl – 39 thous.
  + praca.trojmiasto.pl – 9,9 thous.
* Dedicated data sources (organizations and institutions):
  + www.bazaogloszen.nauka.gov.pl – 1248
  + nabory.kprm.gov.pl – 375
  + BIP – depending on organization

The main conclusion from the analysis of time series is that the fluctuation of job offers among the commercial data sources is high. The differences between 2017 and 2018 were shown in Table 1.

Table 1. Year-by-year comparison in the number of job offers among different data sources

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **No.** | **In Q3 (average)** | |
| 2018 | 2017 |
| **Commercial** | 1 | 64 | 140 |
| 2 | 19 | 124 |
| 3 | 86 | 46 |
| 4 | 54 | 39 |
| **Dedicated** | 1 | 0,6 | 1,2 |
| 2 | 0,8 | 0,4 |
| **Regional** | 1 | 8,5 | 9,9 |

It means that concentration on one data source to acquire and estimate data for the country is not possible. Another issue is that there are lots of duplicates of job offers when taking data from different data sources. So the goal is to select only representative data sources and use the methodological framework to prepare and analyse the data.

The home data were mostly used to prepare the analysis, such as:

* job advertisments by occupation,
* job advertisments by localization,
* job advertisments by occupation and level of education (ISCED),
* trends in job advertisments (by day, by day of week).

The analysis shows that the data from job offers portals are a bit shaky but the scale of different job advertisment in contrast to official job vacancies data is comparable. Such comparison was shown in Table 2.

Table 2. The comparison between official statistics and Big Data sources

|  |  |  |
| --- | --- | --- |
| **SPECIFICATION** | **Data source** | |
| **Official statistics on job vacancies** | **Big Data  (job advertisments)** |
| TOTAL | 122,0 thous. (II quarter 2017)  164,7 thous. (II quarter 2018) | 110,0 thous. (II quarter 2017)  131,0 thous. (II quarter 2018) |
| Mazowieckie voivodship (region) | 27,9 thous. (II quarter 2017) | 23,2 thous. (II quarter 2017) |
| Sales industry | 14,5 thous. (II quarter 2017) | 15,3 thous. (II quarter 2017) |

As shown in Table 2, the differences between official statistics and Big Data for selected data is not a key issue. It means that Big Data source can be used to show the scale of the data, probably not in absolute numbers (due to the existing differences), but rather relative data.

# Methodology of data processing and analysis for home data

The preliminary processing of the web scrapped data was as follows:

* de-duplication of data, i.e. removal of the same job offers, repeated many times,
* manual mapping of the ISCO classification or using machine learning for this purpose,
* matching localities to voivodships in accordance with the TERYT register.

The biggest challenge related to the analysis of job offers is to properly classify the data. Unfortunately, the owners of websites that post job offers allow employers to enter any information in the job offer description. Thus, it is necessary to use text mining methods to verify and assign a job offer to the correct category, expressed by the codes of the Classification of Occupations and Specialties or the International Standard Classification of Occupation (ISCO) codes. Methods supporting the proper mapping of a job offer in the classification include supervised machine learning, which is also not a perfect tool due to the disproportions in the training set in terms of recognized job offers. This means that some job offers may have a greater chance of being properly classified in relation to those less frequent in the training sets. Hence, the experience of the author of the article, as well as other significant centers dealing with this issue, prompts the use of deterministic methods based on regular expressions in the area of ​​text mining, rather than machine learning. Figure 1 shows the differences between the number of job offers before and after the text mining methods mentioned above (orange – before text mining use, yellow – after the use of text mining).

Figure 1. Results of the text mining use for occupation classification

In this case study, for comparison, the training set was prepared manually, i.e. by assigning the name of the job offer to the appropriate code from the Classification of Occupations and Specialties. The methods of analysis used led to the conclusion that it is not possible to obtain the precision of classifying job offers at a level higher than 90%. Hence, the assignment of job offers to occupations on the basis of text mining methods was adopted.

# Main differences between CEDEFOP and home data

The main difference between CEDEFOP and home data analysis is that CEDEFOP data set is collected from various websites, including organizations data, while home data for Poland concentrates on the most representativeness few data sources. One of the most important indicator is how many job offers were scrapped by these two different data sources. It was shown in Table 3.

Table 3. Comparison between CEDEFOP and home data (Big Data - one source)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **COUNTRY** | **year\_grab\_date** | **month\_grab\_date** | **num\_job\_vacancy** | **Home Big Data (one source)** |
| **0** | POLSKA | 2018 | 7 | 100710 | n/a |
| **1** | POLSKA | 2018 | 8 | 65142 | n/a |
| **2** | POLSKA | 2018 | 9 | 147135 | 55821 |
| **3** | POLSKA | 2018 | 10 | 105275 | 55880 |
| **4** | POLSKA | 2018 | 11 | 82683 | 48637 |
| **5** | POLSKA | 2018 | 12 | 62666 | 50251 |
| **6** | POLSKA | 2019 | 1 | 158613 | 44101 |
| **7** | POLSKA | 2019 | 2 | 140211 | 55015 |
| **8** | POLSKA | 2019 | 3 | 112435 | 54241 |
| **9** | POLSKA | 2019 | 4 | 126429 | 52053 |
| **10** | POLSKA | 2019 | 5 | 101187 | 50586 |
| **11** | POLSKA | 2019 | 6 | 106462 | 56843 |
| **12** | POLSKA | 2019 | 7 | 152239 | 54810 |
| **13** | POLSKA | 2019 | 8 | 112736 | 53045 |
| **14** | POLSKA | 2019 | 9 | 170446 | 56002 |
| **15** | POLSKA | 2019 | 10 | 187011 | 57019 |
| **16** | POLSKA | 2019 | 11 | 125873 | 57339 |
| **17** | POLSKA | 2019 | 12 | 99134 | 69515 |

According to the data shown in Table 3, the number of job vacancies based on CEDEFOP is significantly larger than the data based on one data source. Please note that the data from Home Big Data is rather stable, concentrating between 44-69 thous. job advertisments. The data source used for the Home Big Data is one of the most popular and the largest job offer portals in Poland.

In selected months we can observe a weak correlation between data from CEDEFOP and home data. However most of the CEDEFOP data is not correlated with home data sources.

# Main conclusions and future work

The concentration on one data source to acquire and estimate data for the country is not possible as there are huge differences between different job offer portals. The suggestions is to limit number of data sources to the most reliable. For instance, most of the job offers comes from few Internet portals. Webscraping job offers directly from companies websites provides unneccessary redundance and duplicates in the data.

The future work is undergoing in investigations of monitor social skills vs. educational attainment. The preliminary research shows that there is a huge difference between official statistics data and CEDEFOP data. The problem cannot be avoided as there are several mistakes in job offer description – for example, looking for a cleaning personnel is one of the job offers that has high expectations of the skills, for example computer skills and higher education diploma, which is probably a mistake.

# Figures and tables

[Figure 1. Results of the text mining use for occupation classification 4](#_Toc50364694)

[Table 1. Year-by-year comparison in the number of job offers among different data sources 2](#_Toc50364707)

[Table 2. The comparison between official statistics and Big Data sources 3](#_Toc50364708)

[Table 3. Comparison between CEDEFOP and home data (Big Data - one source) 4](#_Toc50364709)

# Literature

Maślankowski J., Pozyskiwanie i analiza danych na temat ofert pracy z wykorzystaniem big data, Wiadomości Statystyczne. The Polish Statistician, vol. 64, 2019, 9, pp. 60-74, DOI 10.5604/01.3001.0013.7590.

Maślankowski J., presentations from workshops/conferences, 2017-2020.