Towards a Common Infrastructure for Online Job Vacancy Data

Pascaline DESCY[*], Vladimir KVETAN[*], Alena ZUKERSTEINOVA[*],
Albrecht WIRTHMANN[+], Fernando REIS[+]

[*]*Cedefop, Thessaloniki,*

[+]*EUROSTAT Task Force on Big Data, Luxembourg, Luxemburg*

Abstract. With increasing penetration of the internet, the number of job vacancies advertised on websites is growing. Although online job vacancies (OJV) are primarily a tool for job matching, the European Centre for the Development of Vocational Training (Cedefop) and the ESSnet Big Data engaged in projects to assess the feasibility of using OJV for labour market analysis and job vacancy statistics. The paper refers to the results of the feasibility studies of both projects identifying opportunities and limitations of OJV for the above mentioned purposes. In addition, it discusses opportunities of creating a joint system for analysing and processing OJVs. During the last two years there have been contacts between Cedefop and the ESSnet Big Data to discuss synergies between both projects. After an initial feasibility study, Cedefop is developing a Pan-EU system providing information on skills demand by 2020. At the same time, the ESSnet will enter into a second phase in November 2018 aiming a creating the conditions for a larger scale implementation of the project, focussing on OJV statistics. This paper outlines a possible partnership between Cedefop and the European Statistical System to create and manage a unique source of OJV data to serve multiple uses in the domain of labour market analysis and official statistics. It presents potential types of (statistical) data and variables based on the information contained in OJVs. Data limitations linked to OJVs nature and specificities will be stressed, too. There is high potential in combining the efforts invested in the Cedefop project and in the ESSnet Big Data to create a joint data collection and processing system on job vacancies that can serve objectives at European, national and regional levels. The paper intends to feed a discussion on the feasibility and the implications of creating a European system for OJV that can serve European and national needs.

1. Introduction

The European Centre for the Development of Vocational Training (Cedefop) is an agency of the European Union. It supports the development and contributes to the implementation of European vocational education and training (VET) policies at EU and national level. In this context, Cedefop is helping policy-makers and other labour market actors to understand supply and demand for labour and skills and any possible imbalances between the two. To complement its various labour market intelligence instruments, Cedefop started developing a pan-EU system for gathering and analysing data contained in online job vacancies (OJVs)[1]. It is foreseen that by 2020, this system will provide information on skills demand from employers in the EU. Cedefop will provide by early 2019 a first early data release for 7 EU countries (CZ, DE, ES, FR, IE, IT and UK). The uniqueness of the data will lie in its ability to provide detailed, timely and frequent information on online job vacancies and skills demand by employers across EU countries.

In the context of the ESSnet Big Data, a group of nine national statistical institutes (NSI) assessed the feasibility of using OJV data in the production of official statistics during a period of 22 months starting in 2016. The group worked on issues of data access and collection, on methodological aspects, such as aligning to the existing definition of a job vacancy, defining a conceptual model for the target population, or extracting information from the vacancy text in natural language[2]. NSI teams identified different possible statistical products and made recommendations for close collaboration between the efforts of statistical offices and the activities of Cedefop.

This paper outlines a possible partnership between Cedefop and the European Statistical System to create and manage a unique source of OJV data to serve multiple uses in the domain of labour market analysis and official statistics. It presents potential types of (statistical) data and variables based on the information contained in OJVs. Data limitations linked to OJVs nature and specificities will be stressed, too. The actors are undertaking various activities to gather data of good quality to ensure that the data fit the needs of the various stakeholders.

There is high potential in combining the efforts invested in the Cedefop project and in the ESSnet Big Data to create a joint data collection and processing system on job vacancies that can serve objectives at European, national and regional level. In its final report[1], the ESSnet recommends that the Cedefop system would form a common infrastructure to support the collection of OJV data for the European Statistical System (ESS) for the second phase of the Big Data project and beyond. The current Cedefop system can be potentially further developed with a view to produce official statistics. This would however require strong inter-institutional collaboration, especially with NSIs,

[1] More information about Cedefop project is available: http://www.cedefop.europa.eu/en/events-and-projects/projects/big-data-analysis-online-vacancies

[2] Cedefop system will be developed to process information in all 24 EU official languages plus other widely used languages in Europe, such as Basque, Catalan, Galician, Gaelic, Russian.

in order to meet the requirements of the European Statistical System, as well as a change in the governance model for the production of European statistics. This will be subject of the final part.

2. What data is gathered and how?

OJVs will be gathered by Cedefop from pre-selected online job boards across all EU Member States. About 500 "sources[3]" (ranging from 50 in Belgium or Netherlands to 3 in Luxembourg) were identified by national experts using quality standards developed in the framework of Cedefop project[3]. Using sources meeting these quality standards allows to build trust in the quality and validity of the data produced. The identified sources represent a significant number of "sites[4]" used to collect information on job vacancies.

Online job vacancies are high dimensional data combining structured and natural language textual data that result from the use of electronic platforms, in particular in the World Wide Web, by labour market operators. It requires the conversion of this digital footprint to relevant statistical variables. The Cedefop project has so far successfully ingested 166 pre-identified sources in seven countries. By the end of August 2018, these sources were connected to more than 38,000 sites. Between April and August 2018, using these sources data pre-processing of over 52 million primary OJVs started.

Preliminary analysis of sources across all EU Member States shows that there are differences between the content of OJVs across web platforms, both between and within countries. Nevertheless, the key common variables that can be classified using the information available from all sources are the following:

- *Detailed occupations* – using the job title and the job description, it is possible to identify the occupational group on the most detailed ISCO level (4 digit).

- *Regions* – the place of work is usually well described in vacancies, allowing classification of the data by region. However, this data should be interpreted taking into account digital divide between regions and countries as well as the industrial composition of regions which will influence hiring practices of local employers.

[3] The term "source" is understoodd as the website / platform /portal that collects information about vacancies to provide these to job seekers.

[4] The term "site" is understood as the actual place on the web, where the final vacancy is posted (for example company website).

- *Skills and other job requirements* – the description of employers' requirements for jobs is the information providing the highest added value for labour market analysis. It includes the following dimensions[5]:
    - *Skills* – although employers rarely advertise jobs using a full job profile, the skills mentioned in OJVs could be considered as critical to assess and select the right applicant for the post. Skills are classified using the ESCO v.1 taxonomy[5][6].
    - *New and emerging skills* – although the majority of skills is expected to be covered by ESCO, the data allows to identify:
        - new skills – skills that do not exist in ESCO or that appear in recent data gathering (such as those required for work with new ICT tools or technologies);
        - diffusing skills – skills appearing in different occupation(s) than the typical one(s);
        - synonyms – new terms used for the same skills.
    - Non-skill requirements – alongside with a relevant skills set, successful candidates are usually requested to meet other requirements (such as having own car, possessing own tools, etc.).
- *Time dimension* – thanks to the high frequency of data gathering, it will be possible to better analyse dynamics of skills requirements within occupations. Moreover, high frequency data can be used for now-casting or short-time projections of skills demand.
- *Other variables* – It is also possible to gather, analyse and present other type of variables contained in the OJVs (e.g. wages, economic activities, type of contract or required experience). This will, however, depend on thorough assessment of the quality and cross country comparability of the data gathered.

3. Potential statistical products

[5] The ability of the system to produce these dimensions was already demonstrated in Cedefop's pilot study.

[6] ESCO is the multilingual classification of European Skills, Competences, Qualifications and Occupations. ESCO is part of the Europe 2020 strategy. For more information please visit: https://ec.europa.eu/esco/portal/home

In its report, the ESSnet Big Data identified potential uses for OJV data. Official statistics on job vacancies are subject to EC regulation No. 453/2008 and are collected primarily for the purposes of calculating the job vacancy rate. The key differences and similarities are summarised in table 1.

Table 1: Differences between the Official survey based and OJV based estimates of job vacancy statistics

Dimension	Official estimates (JVS based)	OJV Data
Frequency	Quarterly	Monthly (or shorter, if required)
Economic activity	Yes	Partial information
Enterprise size	Yes	No
Job title / Occupation / Skills[7]	No	Yes
Sub-national	No	Yes
National totals (estimates)	Yes	Partial information on vacancies advertised online (Yes, when combined with official estimates)

Using OJV data, estimates could be published in higher frequency with shorter time lags between reference and publication date. Due to the selectivity of the OJV data source, they may serve as highly relevant complement to traditional sources in job vacancy statistics. A key advantage of OJV data is that they contain information about job vacancies that are normally not available from traditional sources, yet are often requested by users. This includes classifying OJV data from natural language textual descriptions according to occupation, skills as well as location which could be used to gain insight about the impact of local labour market dynamics in the general employment situation. The activities of the Cedefop project are complimentary to those of the ESSnet in this regard.

The ESSnet recommended to focus on specific problems that would help with the production of experimental statistics, with integration into statistical production being viewed as only a possible longer-term goal.

Analysis of OJV can help improve the statistics produced with data collected via surveys. Firstly, some NSIs have considered the relation between OJV data and official job vacancy estimates to explore the potential of using the near real-time availability of OJV data to produce nowcasts (or

[7] A survey of employer skills for all EU member states was carried on in 2014. Some member states have their own skills surveys, but these are usually infrequent and not at the level of detail possible from OJV.

flash estimates) of job vacancies. A time series approach might be particularly useful for predicting turning points in the economy.

Occupation classification coding frames require regular maintenance to ensure that they capture new job titles. OJV data is an excellent source of information for this. OJV data could be used to support maintenance of the classification itself by reflecting up to date information about job titles and skills in the labour market. There is potential in the use of OJV data to produce new statistics on aspects of the labour market, which are not included in current statistics. One example is international jobs, namely, job vacancies that are advertised in locations that are outside of the country in which the job portal is based. This itself could be a useful measure of labour market tightness and could help identify types of vacancies that are difficult to fill. It may also be that advertisements of this nature are very likely to be advertised only on-line.

4. Data specificity and limitations

While OJVs contain valuable information which can be eventually used to understand vacancy rates and labour market dynamics in ‘real time’, one needs to keep in mind that the primary objective of an OJV is to attract and filter jobseekers to fill an employer’s vacant post. Thus, an OJV is a tool for job matching and not primarily designed for labour market analysis. Although the use of OJVs as a recruitment strategy is growing fast, one need to keep in mind that:

- Not all job vacancies are advertised online as for some types of jobs employers still prefer advertising via more traditional channels (such as newspapers, posters in the windows or word of mouth). It is therefore assumed that OJVs represent only part of labour and job demand and data are subject to an occupational and, to a certain extent, qualification bias.

- In most countries, the online job market is comprised of multiple actors with different business models. Moreover, there is usually no definitive source of all online job ads. Therefore the volume, variety and quality of the data depend on the portals covered by the data collection.

- The actual volume of online job vacancies varies within and across countries and changes over time. Low internet penetration and lack of basic digital skills across the population are the key parameters that influence employers’ strategies to use online portals as recruitment channel.

Processing published online job vacancies to produce sensible data is a new process consisting of various interlinked steps, different from the ones already well known in official statistics production. For instance, it is a common practice that a single vacancy is advertised multiple times on different platforms (or even the same platform) and single advertisements may contain more than one job vacancy. Primary job advertisements therefore need to be de-duplicated[8] and/or expanded[9]. Moreover, variables such as occupation or region need to be derived from the information contained in the job ad. Cedefop uses various automatic algorithms, which need to be programmed and "trained" to deliver good results.

The ESSnet Big Data developed a conceptual model[2] for measuring job vacancies from online sources that describes the coverage of the target population by OJVs and the relationship between job vacancies and OJVs. These models could be applied to improve the Cedefop approach and enlarge the number of statistical products.

The Cedefop system uses various machine learning algorithms for building the automatic classifiers. Regardless of the high precision of the classifiers, data used to train the algorithms may still be affected by processing and classification errors. In addition to setting up a network of experts and establishing quality check mechanisms, Cedefop has entered a lively and fruitful cooperation with Eurostat and the ESSnet Big Data to discuss the rules and procedures (e.g. for classification, de-deduplication, expansion, etc.) aiming at ensuring the highest possible training and output data quality.

5. Addressing data limitations

One needs to stress that the initial and primary objective of Cedefop's project was to produce information on skills demand by employers, which will complement already available labour market intelligence tools. Although innovative and unique, the statistics to be produced will hardly replace traditional labour market statistics. Cedefop's efforts are therefore focussed on presenting and analysing the final data in the context of other labour market data. Furthermore, understanding and managing the specificity of the data is necessary in order to design indicators or carrying out any sort of analysis.

All factors affecting quality need to be understood and addressed. Therefore procedures are developed to provide data of sufficient quality, which are validated by external experts. In order to build trust in the data Cedefop is being transparent and openly communicates the approach and

[8] In the case job advertisement is posted on various places only one job vacancy will be counted.

[9] In the case one job advertisement contain more than one job openings each job opening will be considered one vacancy.

methodology used. This also means that quality issues that are difficult to treat should be documented to allow analysts and end users to use and interpret the data.

There are ongoing discussions as to whether the Cedefop system could be extended for the purpose of producing official statistics and under which conditions. In its conclusions the ESSnet Big Data team working on OJV recommended to reflect on the best use of NSIs time and resources, not underestimating the efforts making OJV data fit for analysis. They recommend considering close collaboration between Cedefop and the ESS to set up a system that would be beneficial to both sides. The current Cedefop project will end in 2020 delivering a fully functional system and related dataset. Significant effort is being put in improving the system by connecting Cedefop with other institutions (Eurostat, ESS-net or universities). This will be a condition to develop the current system into a tool that will be fit for the production of official statistics. However, this goes beyond the purposes of Cedefop's current project and requires substantial future inter-institutional cooperation, with Eurostat and with National Statistical Institutes.

Significant resources would also have to be allocated to this effort. In our view, these are to be focussed on ensuring stable access to sources of primary data, unifying classifiers for jobs and skills developed in Cedefop project and in the ESSnet big data, and assessing and improving data quality.

Stable a*ccess to primary data* is necessary for ensured consistent and robust time series. Currently Cedefop works on the basis of bilateral agreements with owners of various job portals. This implies that the data stream can be affected if portal owners decide to stop the collaboration. Although various strategies are put in place to mitigate dependency on one source, one would need to find sustainable solutions for a system that produces official statistics. Therefore, it would be necessary to ensure closer cooperation between portal owners and NSIs.

Unifying classifiers is crucial for ensuring comparability of final data with various datasets. Cedefop uses ESCO[4] (European Skills/Competence qualification and Occupation) as the key taxonomy for skills and occupations. The occupation pillar of ESCO is fully compatible with the International Standard Classification of Occupations (ISCO)[7], which ensures data comparability across countries. However, the language used by employers in vacancies does not always allow for clear assignment of job titles to ISCO. Cedefop has therefore developed an algorithm to assign job titles and descriptions to ISCO occupations category (ISCO-4 digit). However, this procedure may differ from those used by statistical offices to produce labour force or vacancy statistics.

Assessing data quality is another key prerequisite. Given the unknown though very large size of the total universe of the OJVs, as in other type of big data analysis, testing is made on data set which represents the best available data under reasonable conditions for training and testing the machine learning algorithms. Based on Cedefop estimations, one would have to test manually about hundred thousand vacancies per language pipeline to achieve good and reliable results (60% of data is used for machine learning, 20% for testing and 20% to evaluate selected model). Taking into consideration, that Cedefop is developing language pipeline for each EU official language and few more significant languages across EU (for example Catalan, Russian or Norwegian) the costs of such testing is far beyond the current possibilities of Cedefop and would also require inter-institutional cooperation.

6. Conclusions

Cedefops efforts are focused on analysing labour market trends to understand current and future skills needs within the European Union. These efforts can be complementary to the objectives of the ESS to improve job vacancy statistics in various aspects. In addition, skills of both parties could complement, too. On the side of official statistics this comprises the conceptual approach, the link with existing data, application of statistical methods and the commitment to quality.

By the end of 2020, a system producing unique dataset will become available, using big data technologies to gather detailed information on OJVs and skills demand by employers across EU countries. This dataset will meet Cedefop needs for new and evolving data on skills needs, which will usefully complement traditional labour market data produced for all EU countries. It is in Cedefop’s intention to develop the system further and continue with gathering and analysing the data also beyond 2020.The system could also serve needs of the ESS to improve quantitative statistical information on the labour market. The related production system could be adjusted to serve as unique source of information at European, national and regional level. The system could be further improved by ingesting knowledge on the online labour market at national level, e.g. for training classifiers. This system would provide economies of scale running one instead of 28 systems and by processing OJVs by language pipelines. The production of European statistics based on a common system, including data capture, pre-processing and information extraction, would provide highly harmonised statistics. The system would serve multiple purposes and users, providing pre-processed data for further analysis or ready to use information for end users. Agreements with multinational OJV platforms at European level would contribute to the economies of scale. There is an opportunity to further develop and expand the system for collecting, processing OJVs for the production of official statistics. However, this requires strong inter-institutional cooperation as well as a new model of statistical governance. With the contribution of Cedefop and the ESSnet Big Data, Eurostat will start developing a business case outlining different scenarios of integration that will serve as basis for a decision on realising a common system of OJV data in 2019 that will be able to serve different users at European and national levels.

7. References

[1] ESSnetBig Data (2018), Web Scraping / Job vacancies, Final technical report (SGA2), https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5e/SGA2_WP1_Deliverable_2_2_main_report_with_annexes_final.pdf.

[2] ESSnet Big Data (2017), Web scraping / Job vacancies, Interim Technical Report (SGA1), https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/64/WP1_Deliverable_1_2_final.pdf

[3] Kvetan, V. (2018), Understanding the online labour market in the EU, https://skillspanorama.cedefop.europa.eu/en/blog/understanding-online-labour-market-eu

[4] M. le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandensteen and J. De Smedt, "ESCO: Boosting Job Matching in Europe with Semantic Interoperability," in Computer, vol. 47, no. 10, pp. 57-64, Oct. 2014.

[5] European Commission (2013), ESCO: European Classification of Skills/Competences, Qualifications and Occupations, http://bookshop.europa.eu/en/esco-european-classification-of-skills-competences-qualifications-and-occupations-pbKE0313496.

[6] de Smedt, J., le Vrang, M., Papantoniou, A. (2015), ESCO: Towards a Semantic Web for the European Labour Market, www2015 workshop: Linked Data on the Web (LDOW2015).

[7] International Labour Office (2012), International Standard classification of occupations, ISCO-08, Geneva.