Jasper Van Loo

Big data, better skills intelligence?
Using online job advertisements to map skills trends

Jasper Van Loo[1], Vladimir Kvetan[1]

[1] Cedefop – European Centre for the Development of Vocational Training, Thessaloniki, Greece, Jasper.van-Loo@cedefop.europa.eu; Vladimir.Kvetan@cedefop.europa.eu

Abstract:

The digital transformation and shift towards a climate-neutral Europe are changing how we work, learn and take part in society. Countries and communities can only grasp their opportunities if people develop the right skills and have opportunities to use them. To bridge the worlds of work and education the European Centre for the Development of Vocational Training (Cedefop) has produced skills intelligence for the European Union for over a decade. Skills intelligence is the outcome of an expert-driven process of identifying, analysing, synthesising and presenting quantitative and/or qualitative skills and labour market information. These may be drawn from multiple sources and adjusted to the needs of different users.

In rapidly changing labour markets skills intelligence needs to be based on fast (ideally real-time) and disaggregated (at the level of region) information. Moreover, it needs to allow for the identification of actual skills beyond commonly used skills proxies such as occupations or formal qualifications. While conventional statistics such as surveys remain an important and relevant source, using them to dynamically map change in the world of work in a great level of detail is challenging.

Information on skills available on the internet has the potential to enrich labour market analysis and skills intelligence and is currently high on the agenda. Although booming in social science research, the use of such data for developing policy-relevant skills intelligence is still in the early stages. In the last few years, Cedefop has been developing a pan-European system for gathering and analysing data from online job advertisements. The system collects job advertisements from online job portals and websites in all 27 EU Member States and the United Kingdom and processes the text contained in them in all EU official languages.

This paper presents key features of Cedefop's big data production system, which is part of Eurostat's Web intelligence hub, and gives an overview of its future potential and use cases. It aims to develop understanding of various biases and provide guidance on correctly interpreting findings as this is critical for drawing accurate conclusions and developing suitable policy recommendations. While big data offers interesting opportunities for skills intelligence, it by no means can replace conventional information. Surveys and administrative data remain relevant, not only in their own right, but also because they can validate big data analyses and provide the context necessary to ensure findings are well understood and interpreted.

Keywords:
Web intelligence, Labour market

1. Introduction:

In rapidly changing labour markets skills intelligence needs to be based on fast (ideally real-time) and disaggregated (at the level of region) information. Moreover, it needs to allow for the identification of actual skills beyond commonly used skills proxies such as occupations or formal qualifications. While conventional statistics such as surveys remain an important and relevant source, using them to dynamically map change in the world of work in a great level of detail is challenging.

To collect such information, in the last few years, Cedefop has been developing a pan-European system for gathering and analysing data from online job advertisements (OJA). The system collects OJA from online job portals and websites in all 27 EU Member States and the United Kingdom and processes the text contained in them in all EU official languages. The first part of this paper presents key features of Cedefop's big data production system (DPS) which is part of Eurostat's Web intelligence hub (WIH). The second part showcases some key results and gives an overview of how these can enrich LMSI. The final part aims to develop understanding of various biases and provide guidance on correctly interpreting findings. This is critical for drawing accurate conclusions and developing suitable policy recommendations. This paper builds primarily on two publications related to the topic: Cedefop (2019) and Cedefop et al. (2021).

2. Methodology: Key features of Cedefop's OJA data production system

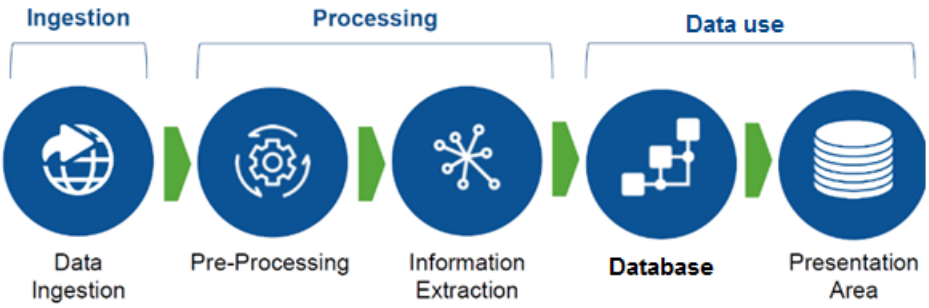
Producing web-based big data requires a data production system (DPS) for data ingestion, data pre-processing, information extraction and data use/presentation (Cedefop, 2019). Cedefop's refers to the information collected through its DPS as online job advertisements (OJA). This term is most suitable to be used when referring to online documents containing information on current vacancies ([1]). As Cedefop's DPS operates in the multilingual environment of the EU ([2]) it is organised as a set of individual language pipelines allowing to process the data in the original language prevents information losses by automatic translations. Each language pipeline is based on a modular system of individual activities (see figure 1).

Data ingestion refers to the process of gathering primary documents from the web. This process is usually also referred to as "web scraping" and "web crawling". Although often these terms are used as synonyms, in practice there are significant differences between these activities. In addition to the above two data ingestion methods, a general preference is to ideally reach an agreement with the data owners to organise. (for more information about data ingestion see Cedefop, 2019 pp 18-21).

([1]) This term is preferable over "Online job vacancy" because a vacancy is defined as "a job opening, which is newly created, unoccupied or about to become vacant and for which an employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned. The employer intends to fill the opening either immediately or within a specific period of time." (See Eurostat glossary: https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Job_vacancy_rate_(JVR)) There can be a wide variety of "active steps of an employer" and channels to advertise a job opening (newspapers, banners in window, online). Some do not advertise it at all. Smaller employers often resort to informal networks or social contacts.

([2]) There are 23 official languages in the European Union. Moreover, in some countries OJAs are published in other languages officially recognised in the country (for example Catalonian or Basque in Spain). It is also wide practice that in one country OJAs could be published in other language (for example in Slovakia there could be OJAs published in German, English or even Hungarian).

Figure 1. **Web based data collection and production process**

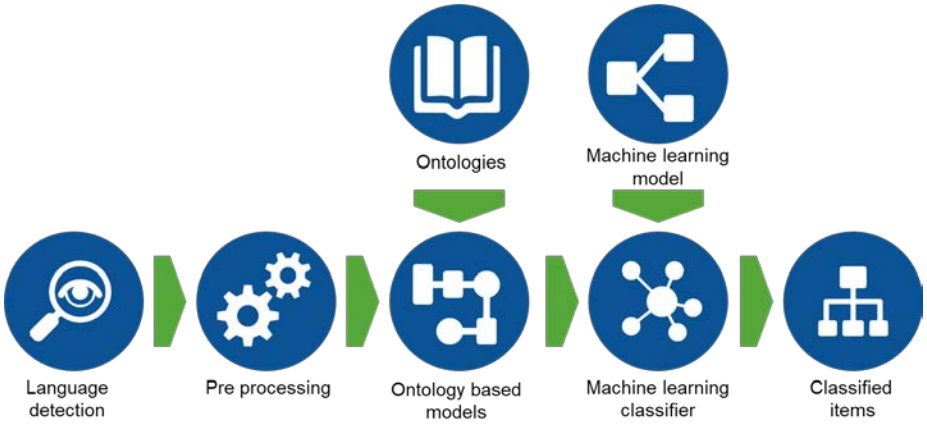


Source: Cedefop (2019)

When working with multiple web-based OJA sources (public and private jobportals, corporate websites etc), they can naturally differ in quality and content. To develop a database suitable for subsequent analysis, data pre-processing is needed. This involves cleaning (deleting all non OJA content), merging (enriching information from OJA published in multiple sources) and de-duplicating (discarding OJAs published in multiple sources).

The pre-processed data are used in the information extraction process. This step in the process has two important features: ontologies and machine learning models (Figure 2). Ontologies (ideally multilingual) create a framework for processing and analysis of documents. Cedefop's DPS uses standard ([3]) and custom ([4]) ontologies.

Figure 2. **Information extraction process**



Source: Cedefop (2019)

With the power and flexibility of machine learning algorithms, the contents of extracted documents are subsequently matched to available ontology terms, such as occupation,

([3]) Standard ontologies refer to well set classifications maintained by external organisation, such as ISCO for occupations, ESCO for skills, NACE for industry, NUTS for geographical unit, ISCED for educational level.

([4]) Developed on the basis of information available in documents, for example type of contract, experience, salary or working hours in OJA data

industry, region of the workplace or type of contract. To do that, a DPS can employ different strategies – exact text matching, text similarity or machine-learning algorithms.

Ontologies should be continuously updated and enriched. The results of the machine classification process should ideally be regularly validated by domain and language experts. The outcomes of this process (proposed corrections) should subsequently be used to further improve the accuracy of the machine classification or to enrich ontologies. This semi-automatic augmentation process can add new terms and synonyms not yet included in the ontology through machine learning algorithms after approval by experts. Ontologies can also be updated manually, to reflect new information (new trends in occupations or even updating the whole ontology such as the European Skills, Competences, Qualifications and Occupations (ESCO) classification([5])).

After data processing, the data are stored in a multidimensional database. Cedefop's OJA database is accessible through two channels. The first is the data presentation platform (DPP) SkillsOVATE ([6]). This online platform features user-friendly visualisations allowing non-technical users to navigate through the data without need for high-level programming skills. The other channel is a data lab, which allows users to perform basic and advanced data science analysis using processed data stored in the database.

3. Results: Showcasing the potential of OJA data

The descriptions of what skills employers are looking for in job advertisements offer the highest added value for labour market and skills analysis. Using web-based, human-sourced, documents to better understand skill demand and supply has several advantages compared to skills information collected via traditional statistical methods, such as surveys. There are, for example, clear limits to using an employer survey to understand skill needs and trends in occupations; only a limited number of skills can be considered; simplification is needed to keep the questionnaire manageable for respondents and it is difficult to capture emerging skill needs systematically. And unless the survey is large (and costly), analysis typically remains at an aggregated level to derive reliable findings. Cedefop; European Commission; ETF; ILO; OECD; UNESCO (2021).

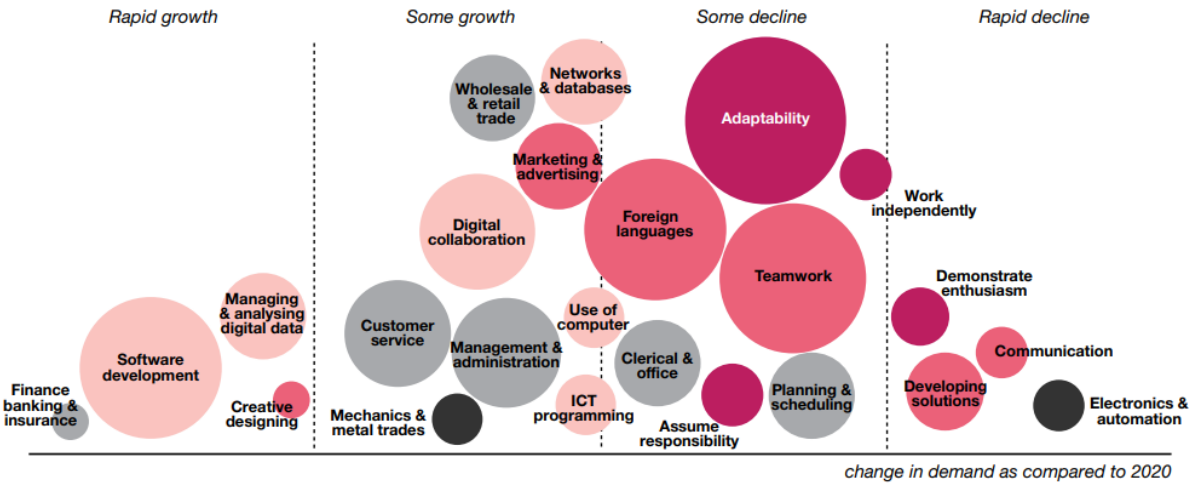
Online job advertisements (OJAs) offer possibilities to obtain more detailed information on trends in jobs and skills. Cedefop's Skills OVATE OJA analysis and presentation tool presents occupations at the three digit level of the International Standard Classification of Occupations (ISCO[7]) while data for expert users the four digit level is available. Regional analysis going beyond what is possible with many conventional sources is possible because most employers provide detailed information on the place of work in the description of jobs they advertise online. In case a sufficient number of reliable observations is available, the information can be also facilitate labour market and skills analysis at the local level (e.g. in cities and metropolitan areas).

As OJA are high frequency data, undertaking trend analysis, nowcasting and production of short-term forecasts is possible. It also enables analysing labour market developments and trends in skills requirements within occupations over time. Such analysis was used by Cedefop to estimate effects of COVID 19 on the labour market Cedefop (2021). While labour markets and hiring were heavily hit by the outbreak of pandemic, the first signs of improved situation

([5]) European Skills, Competences, Qualifications and Occupations (ESCO) https://ec.europa.eu/esco/portal/home
([6]) Accessible via https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies
([7]) International Standard Classification of Occupations (ISCO) https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/

reflected in relatively quick recovery. Cedefop's OJA analysis helped uncover that the manufacturing sector drove the recovery and showed that apart from digital skills, in particular skills linked to creativity and adapting business models and sales were trending in 2020 (Figure 3). The possibilities to look at digital skills in detail makes it possible to gain insight into the consequences of to the digital transition speeding up and the re- or upskilling needs linked to it. While some digital skills (e.g. digital collaboration – such as via virtual meetings) are relatively easy to learn, others (programming, software development) represent much more fundamental learning needs. Understanding in which labour market segments (e.g. occupations, sectors, regions) basic, intermediate and advanced skills needs are emerging is of great value for public policies and corporate HR strategies.

Figure 3. **Skills trending in 2020**



Source: Cedefop (2021)

4. Discussion and conclusions: Understanding potential and limitations of OJA data

Although there is no doubt about the potential of OJA data producing valuable and interesting information on labour market and skills trends, there several caveats. As the nature of such data is fundamentally different compared to data collected via traditional methods, also the challenges are of a different nature. Big data pose marked challenges in terms of the representativeness of the information collected and is often fraught with sample selection problems (Cedefop et al., 2021). This makes it challenging to apply traditional statistical techniques and tests and to draw generalised conclusions about underlying populations.

There is selection bias because not all vacancies are advertised online and not all job advertisements lead to actual job openings. Employers use different hiring strategies for different occupations. Vacancies for high level professionals are often advertised on dedicated or privately-owned portals, or job-hunting is used to fill them. In many countries, medium- or lower-level posts are typically advertised on public employment service (PES) portals. There are also jobs that are rarely advertised online at all, because word of mouth or a notice in a shop window are more effective and cheaper solutions to filling vacant posts. Moreover, some web portals target particular types of users, as is the case with some PES portals which are accessible only to the registered unemployed.

Other conceptual challenges relate to the need to analyse big data in the context they originate. Language, culture and how the labour market and education systems are organised and function affect the shape and content of information published online. In countries where

professions are heavily regulated, for example, OJAs are not as rich in terms of skills as is the case in countries with more liberal systems ([8]). This affects cross-country comparability of analyses based on big data and complicates using multilingual taxonomies for skills and occupations. Digital divides and the economic structure must also be considered when comparing results between countries and within countries and regions.

Providing reliable and meaningful insight requires also sophisticated methods and tools. Artificial intelligence (AI) methods, such as natural language processing (NLP) algorithms, are used to better understand and interpret new terms appearing in OJAs. Machines take over a lot of work, but the 'human intervention' in terms of training, validating and interpreting results is and will remain essential. Many documents must be classified and then checked manually, to train and improve the machine learning algorithms. Big data analysis, despite all its AI glamour, is resource intensive and requires a variety of expertise. It is not possible to rely on computational and algorithmic power alone. It is the combination of artificial intelligence and human intelligence that is key for producing and interpreting web-based big data that supports skills policies.

Despite these caveats we believe OJA data analysis is a promising way forward to enhance or complement conventional labour market and skills data and – in a longer term perspective – also to compile new statistics. Cedefop cooperates closely with the network of the European Statistical System (ESSnet) and Eurostat to set priorities and shape conceptual and developmental work. The technical complexity of the work and the need to involve a broad circle of stakeholders has made it clear that a more coordinated approach is needed to move towards using OJA for official statistics (Descy et al.,2019). A major step towards more coordination was moving the collection and analysis of OJA data to Eurostat's Web intelligence hub in 2021.

References

1. Cedefop (2019), Online job vacancies and skills analysis: a Cedefop pan-European approach. Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/097022

2. Cedefop; European Commission; ETF; ILO; OECD; UNESCO (2021). Perspectives on policy and practice: tapping into the potential of big data for skills policy. Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/25160

3.Cedefop (2021): Trends, transitions and transformation, Cedefop briefing note, https://www.cedefop.europa.eu/en/publications-and-resources/publications/9157

4. Descy et all (2019) Descy, P., Kvetan, V., Wirthmann, A., Reis, F., Towards a shared infrastructure for online job advertisement data, Statistical Journal of the IAOS, vol. 35, no. 4, pp. 669-675, 2019
. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190547

([8]) As an example, while in countries with highly regulated professions a vacancy saying "plumber needed" without elaborating the skills required suffices, in countries where this is not the case, employers may need to list the skills requirements in such a vacancy in detail to filter out appropriately trained or skilled candidates.