Perspectives on policy and practice

Tapping into the potential of big data for skills policy

CEDEFOP
European Centre for the Development of Vocational Training

ETF
Working together Learning for life
European Training Foundation

ILO
International Labour Organization

European Commission

OECD
BETTER POLICIES FOR BETTER LIVES

UNESCO
United Nations Educational, Scientific and Cultural Organization

CEDEFOP

Perspectives on policy and practice

Tapping into the potential of big data for skills policy

Luxembourg: Publications Office of the European Union, 2021

Please cite this publication as:
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

Luxembourg:
Publications Office of the European Union, 2021

PDF
ISBN 978-92-896-3235-5
doi:10.2801/25160
TI-09-21-027-EN-N

Designed by Missing Element Prague

The **European Centre for the Development of Vocational Training** (Cedefop) is the European Union's reference centre for vocational education and training, skills and qualifications. We provide information, research, analyses and evidence on vocational education and training, skills and qualifications for policy-making in the EU Member States.

Cedefop was originally established in 1975 by Council Regulation (EEC) No 337/75. This decision was repealed in 2019 by Regulation (EU) 2019/128 establishing Cedefop as a Union Agency with a renewed mandate.

Europe 123, Thessaloniki (Pylea), GREECE
Postal address: Cedefop service post, 57001 Thermi, GREECE
Tel. +30 2310490111, Fax +30 2310490020
Email: info@cedefop.europa.eu
www.cedefop.europa.eu

Jürgen Siebel, *Executive Director*
Barbara Dorn, *Chair of the Management Board*

Contents

Acknowledgements

This perspective on policy and practice was led by the European Centre for the Development of Vocational Training (Cedefop) on behalf of the inter-agency TVET working group on Skill mismatch in digitised labour markets. Jiri Branka, Vladimir Kvetan, Konstantinos Pouliakas and Jasper van Loo (Cedefop) coordinated the work and took the lead in drafting it. The ETF (Anastasia Fetsi, Francesca Rosso and Eduarda Castel Branco), OECD (Luca Marcolin, Marieke Vandeweyer), ILO (Bolormaa Tumurchudur-Klok, Ana Podjanin, Olga Strietska-Ilina), UNESCO (Hiromichi Katayama, under the guidance of Borhene Chakroun and Hervé Huot-Marchand) and the European Commission (Michael Horgan) contributed parts of the text and provided examples of global practice in the field of using big data for skills analysis.

Key messages

- Policy-makers need faster and more detailed information on skills to monitor and respond to the challenges created by structural economic and societal megatrends and the Covid-19 pandemic.
- Providing information in (quasi) real-time, online labour market data have great potential to improve policy-makers' understanding of trends in skills needs and supply.
- The strengths of web-based big data include timeliness and granularity compared to conventional approaches to skills analysis.
- While web-based big data have significant potential for skills policy, they tend to require more effort to prepare for analysis than data collected using conventional approaches. The unstructured information provided often suffers from statistical, selection and conceptual biases.
- In low-income countries, web-based big data analysis can provide useful insights that complement conventional skills analysis, but biases can be more challenging. Higher informal employment and a less-developed digital infrastructure means online recruitment covers only a small part of the job market, particularly urban, formal and white-collar jobs. This complicates analysis that aims to cover the wider labour market.
- Despite advancements in information and natural language processing (NLP) and cloud computing, setting up a stable and well-functioning system for gathering, processing and analysing big data remains challenging. Developing such a system is a complicated and resource-intensive endeavour, but one that can pay off in the long-run.
- Web-based big data cannot and should not replace other skills intelligence methods and sources. Exploiting the complementarities of big data and other sources of skills intelligence is key in generating statistically robust, detailed, and policy-relevant evidence.
- It is the combination of artificial and human intelligence that will be key for further developing big data's role in shaping effective technical and vocational education and training (TVET) and skills policies in the coming years.

CHAPTER 1.

Introduction

Demographic change, the shift towards more sustainable economies, digitalisation and new forms of ICT-based work are reshaping skill supply and demand around the world with wide-ranging economic and societal consequences. These trends, which may have been accentuated by the Covid-19 pandemic, are profoundly affecting the labour market and increasing the uncertainty around future skill needs. To shape effective skills policies, decision-makers need faster and more detailed collection and analysis of information on current and future skill needs and trends.

Using information available online – or 'web-based big data' (Box 1) – for labour market analysis and skills intelligence is currently high on the policy agenda. While big data analysis is booming in social science research, its widespread use for labour market or education and training policies is still limited. The main reason lies in the very nature and requirements of developing and using such data.

Box 1. **What are web-based big data?**

Big data are not simply a large data set. The population census of India, which has more than 1.3 billion records, is still considered conventional data, because it is collected using standard methods. 3 Vs – high-volume (amount of data), high-velocity (speed of generation and collection, rendering it almost 'real time') and/or high-variety (range of different data types and sources) – make data 'big data' (Laney, 2001). Experts in the field have proposed (ETF, 2019) two additional Vs: veracity (accuracy and data quality, given that quality cannot be controlled at source) and value (extent to which stakeholder information needs are met).

Most big data belong to one of three types:

- human-generated (individuals submitting own information in social networks, web platforms);
- process-generated (credit card data, financial transactions);
- machine-generated (data collected via sensors, mobile phones, internet of things).

In the context of labour market and skills policies, human-generated information available online is most important. The internet can be used to assess and analyse skills supplied and demanded in job markets. Commonly used sources include electronic CVs available through online platforms or social networks, job advertisements published on job portals, and online descriptions of education and training programmes and qualifications on offer. In this publication, we use the umbrella term 'web-based big data' to refer to all these sources.

Source: Inter-agency technical and vocational education and training (IAG-TVET) working group on Skill mismatch in digitised labour markets. (hereafter IAG-TVET working group).

Web-based big data are based on sources that are not primarily designed for labour market and skills analysis. Firms post job advertisements to attract the best candidates to their vacant posts. Jobseekers interact with web platforms and tools to showcase their skills and potential to prospective employers. The information published by education and training institutions and government agencies responsible for regulating programmes and qualifications also represents a wealth of skills-centred big data.

Experts in charge of designing systems for gathering and analysing web-based big data are not fully in control of the information generation and collection process. The external sources used usually have uneven coverage across occupations, job types, skills levels, sectors and countries. As a result, unprocessed big data raise particular challenges for analysis. Developing a comprehensive and robust system for data collection and analysis which can mitigate such challenges is a complicated and costly endeavour. Policy-makers interested in developing such a system should factor in substantial initial investment and be aware of the costs involved in keeping it operational.

This publication has been prepared by the interagency TVET group on Skill mismatch in digitised labour markets, to support experts and policy-makers who wish to engage in discussion on the potential of web-based big data for skills policy. It outlines how such data can be used to mitigate labour market challenges, reduce skills mismatches and strengthen the links between the labour market and education and training. The focus is on overcoming conceptual and practical challenges and limitations, system development and using big data for skills policy in practice. Examples of big data initiatives from around the globe illustrate its potential and provide insight into how big data are already supporting policy-makers in shaping the futures of work and education.

CHAPTER 2.

Labour market and skills trends: the increasing importance of big data

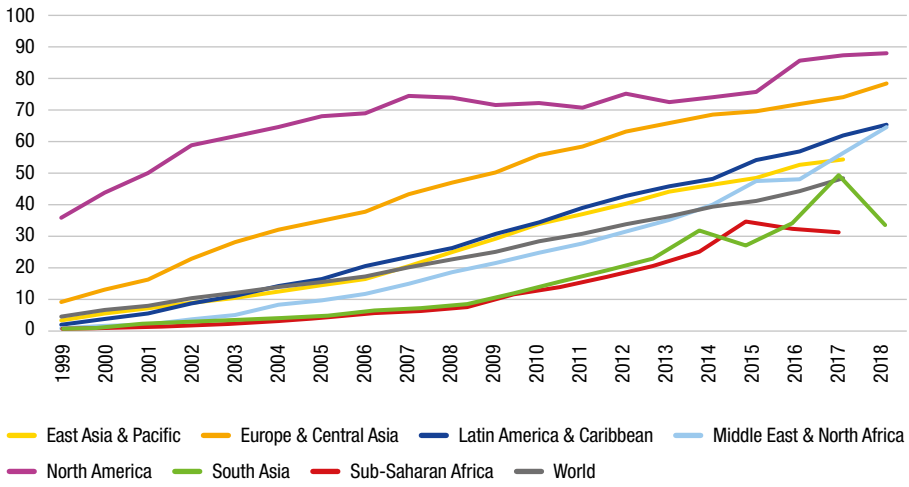
Using web-based big data for skills analysis requires a mature online job market. A well-developed internet infrastructure and good and widespread connectivity are preconditions. Without them – as is the case in many low-income countries – the online job market will remain marginal, as many individuals and employers will not be able effectively to communicate and interact online. Collecting web-based big data in such a context is unlikely to result in information that can be used for solid skills and labour market analysis.

While internet access in sub-Saharan Africa and South Asia remains limited, with internet availability steadily improving and converging across the globe since 2000 (Figure 1), the potential of collecting and using big data and the demand for it will continue to grow in coming years. Online recruitment and job search are becoming more important. In 2018, applications received through online job portals accounted for a fifth of hires worldwide.

Digitalisation, growing internet penetration and increasing digital literacy of the population directly drive the use of the web as a labour market and education and training intermediary (Cedefop, 2019a). The economic situation, the structure of labour demand and supply, and the digital preparedness and engagement of education and training institutions, employers and public employment services are important indirect factors. The degree of mismatch in the economy also plays a role. Skill shortages incentivise employers to rely more on online job portals; high skill underutilisation and underemployment may lead individuals to search more actively for job and education and training opportunities online. Legislation or regulation mandating the use of online tools is also an important factor contributing to the proliferation of the internet in labour market and education and training settings ([1]).

([1]) In some countries all vacancies must be advertised via a public employment service (PES) website. The unemployed also have to be registered and post their CVs in dedicated web

Figure 1. **Individuals using the internet in different parts of the world (% of population)**



Source: International Telecommunication Union (ITU) World telecommunication/ICT indicators database.

The expansion of online information on jobs and skills has value going beyond the direct benefits users derive from it. Web-based, human-sourced big data can play a key role in developing policy-relevant skills intelligence. Apart from helping employers in finding talent, online job advertisements (OJAs) can also be analysed with a view to uncovering sectoral, occupational and skills trends. On top of their role as tools to promote individuals to prospective employers, online CVs and personal social media profiles can be analysed to obtain insights into jobseekers' skills and work experience, career paths and mobility, and engagement in training and learning.

Information that characterises education and training programmes and their outcomes, such as programme descriptions, curricula, learning outcomes/skills and qualifications, can give insight into gaps between education and training provision and skills needs. Electronic patent and scientific paper repositories can be analysed to understand better the skill needs arising from the diffusion and adoption of technologies. Such

platforms; schools and training institutions are sometimes obliged to post their study programmes and training offers online. See Cedefop (2019a).

technology-focused skills analysis makes it possible to look ahead and identify leading skills trends which may not yet be visible ([2]).

This publication focuses on using web-based labour market data contained in online job advertisements and CVs for skills analysis. Advancements in information and natural language processing (NLP) and cloud computing have vitally contributed to the development of big data analysis for skills. Web-based big data on skills can be used to generate evidence that complements other types of skills intelligence, such as skills forecasts, analysis based on surveys or administrative data ([3]). It can support policy-makers and governments in developing more focused and customised skills policy interventions. Notwithstanding its potential, providing useful and novel insights, generating policy-relevant, reliable and high-quality statistical data using big data is not straightforward. To gain experience and insight into how to address key challenges, international organisations have taken a leading role in developing approaches and piloting systems using web-based big data (Box 2).

Box 2. **Unlocking the potential of web-based data for skills analysis: the role of international organisations**

Cedefop has developed Skills OVATE, a system for gathering and analysing online job advertisements (OJAs) in the European Union. The project will be further developed in cooperation with Eurostat, to pave the way for producing official labour market statistics. Cedefop's OJA project can support other EU initiatives, such as the European taxonomy for skills, competences and occupations (ESCO). Analysing information provided by Europass CV users, the agency has also piloted big data analysis of skills supply. Cedefop's big data work in the coming years will focus on providing evidence in support of the up- and reskilling ambitions put forward in the 2020 EU skills agenda.

Building on the achievements of the ESSnet big data project and the long-term cooperation with Cedefop, Eurostat is moving towards tapping the potential of big data to feed into official statistics by implementing the agreements in the Scheveningen memorandum. To serve EU and other international institutions, it is developing the

([2]) More information on using patent and bibliometric analysis can be found in: Cedefop (2021, forthcoming). *Guide on methods and practices of anticipating new technologies and skills.* The OECD has a long history of analysis of patent and bibliometric data. See, for instance the OECD STI scoreboard platform and *Measuring the digital transformation* (OECD, 2019b).

([3]) See compendium of six *Guides on skills anticipation methods* produced by ETF-Cedefop-ILO.

Web Intelligence Hub, a big data infrastructure which will become a central access point for various types of information.

In 2019 the ETF started work on big data for labour market intelligence (LMI) and produced a methodological big data guide and a feasibility study with guidelines for users. The aim of this new area of work is to explore the potential of data analytics to improve the performance of conventional LMI in ETF partner countries (transition and developing countries surrounding the EU). The scope of the analysis is skills demand. The initiative blended exploratory work in mapping the conceptual and methodological underpinnings developed in different countries and research projects. Following the feasibility phase based on landscaping sources of online job advertisements in 2019 ([4]), in 2020 the ETF started developing online job advertisement collection and analysis systems for Tunisia and Ukraine. In parallel, the ETF launched several studies on the future of skills in economic sectors and used big data to complement other empirical research methods. Text mining techniques were applied to collate data on emerging technological trends from patent data and scientific papers, and to identify emerging skills needs associated with them. The ETFs big data initiative also comprises technical dissemination actions and training of statistical and analytical departments and experts, and contributes to the ETF Skills Lab.

The ILO has used online job advertisements to assess skills needs. In the context of its study on a transition to an environmentally sustainable economy, model-based work was combined with US OJA data provided by Burning Glass Technologies (BGT). The OJA data was used to proxy employer skills requirements in order to understand their reskilling needs ([5]). The same BGT data set was used in a study analysing the change in skills demand in the context of global trade. The 2020 ILO report The feasibility of using big data in anticipating and matching skills needs bundles contributions from participants to a 2019 ILO workshop on the topic. A pilot study to develop a methodology for defining a skills framework for the Uruguayan labour market based on job advertisement and job applicant data was in progress at the time of writing this publication.

The OECD is leveraging several sources of big data to support policy analysis and recommendations. Its 2015 recommendation on good statistical practice advocates that national statistical offices explore internet-based sources, and the combination of these with existing sources for official statistics. In the areas of employment, social affairs and education, data on hiring and online job vacancies are used to an-

([4]) A landscaping study of the online labour market and ranking of OJA sources has also been conducted for Belarus.

([5]) The results of this study are presented in the ILO global report: ILO (2019). *Skills for a greener future: a global view based on 32 country studies*.

alyse online residual labour and skills demand, describe the career paths of tertiary graduates, investigate patterns of diffusion of digital or AI technologies and their consequences on the labour market, and improve business cycle forecasting ([6]). During the Covid pandemic, these data enabled timely analysis of labour market dynamics, including the differential impact of the pandemic on labour market demand across US cities. The next update of the OECD Skills for jobs database will include a module based on OJA data to strengthen its measurement of skills imbalances. The OECD AI Policy Observatory gathers and presents information on, among others, labour market policies related to the diffusion of artificial intelligence. It also offers data visualisations of selected web-based labour market big data.

UNESCO uses TVET and labour market data to identify and anticipate trends to inform Member States about the future of skills supply and demand in the labour market within the framework of its TVET strategy. It also supports the development of data-backed policy and programmes. Due to the cross-cutting nature of TVET and fragmentation of data and statistics, and the lack of data integration between different Ministries and the private sector, it is difficult to capture accurately the status of skills supply and demand in the labour market, which is critical for TVET policy development and implementation. While traditional LMI, including administrative and survey data, already offers a detailed picture of the status of labour markets, big data can help improve it. UNESCO's experience in Malawi and Myanmar demonstrate the potential for combining traditional LMI with big data from online job-search platforms.

Source: IAG-TVET working group.

([6]) See:
OECD (2017). *Digital economy outlook*.
OECD (2019a). *Benchmarking higher education system performance*.
OECD (2019b). *Measuring the digital transformation: a roadmap for the future*.
OECD (2020a). *OECD Employment outlook 2020: worker security and the COVID-19 Crisis*.
OECD (2020b). *Skills measures to mobilise the workforce during the COVID-19 Crisis – OECD Policy responses to coronavirus*.
OECD (2020c). *Labour market relevance and outcomes of higher education in four US States: Ohio, Texas, Virginia and Washington*.
OECD (forthcoming). *Measuring the impact of the COVID-19 crisis on jobs and skills demand*.
OECD Policy responses to coronavirus.
OECD (forthcoming). *OECD Skills outlook 2021*.

CHAPTER 3.

The value added of big data for skills analysis

Using web-based, human-sourced documents to understand skill demand and supply better has several advantages compared to skills information collected via conventional methods, such as surveys and administrative data. There are clear limits to using an employer survey to understand skill needs and trends in (detailed) occupations, as only a limited number of skills can be considered; simplification is needed to keep the questionnaire manageable for respondents and it is difficult systematically to capture emerging skill needs. Unless the sample of the survey is representative (large and costly), analysis typically remains at aggregate level to derive reliable findings.

Producing web-based big data requires a data production system (DPS) for data ingestion, data pre-processing, information extraction and data use/presentation (Cedefop, 2019b); see Box 3. While developing such a system is complex, the data it provides allow for greater precision in estimates thanks to the large number of observations available and information granularity. This contrasts with well-designed survey data sets, which typically provide unbiased estimates of population parameters but often with lower precision. The main types of information that can be extracted from online job advertisements and online CVs (Figure 2) can be used to analyse:

(a) skills demand and supply patterns: although employers rarely use a complete skills profile to advertise jobs, the skills (proxies) mentioned in OJAs to assess and select the right applicant for the post provide detailed insight into skill needs in occupations and sectors. Such information is difficult if not impossible to obtain via other means ([7]). CVs, in which individuals increasingly emphasise their job-specific and transversal skills (such as language and ICT skills) on top of their formal qualifications and work experience, can help in characterising elements of skills supply. The

([7]) For example, the detailed list of 'skills terms' in OJAs can be classified using standard taxonomies, such as the European skills, competences, qualifications and occupations taxonomy (ESCO v.1) or the O*NET taxonomy.

key job tasks and responsibilities people list in their CVs can be used to shed light on job complexity;

(b) new and emerging skills: OJAs help identify emerging skill needs linked to new tasks and technologies, such as those that are not part of standard taxonomies (e.g. ESCO, O*NET);

(c) skills at regional or local level: OJAs typically describe the place of work well, which facilitates analysis at regional and – provided a sufficient number of observations is available – local level. Information on skills demand, supply and trends at these levels can be used to strengthen skills ecosystems;

(d) diffusion of skill requirements: OJAs can be used to map the proliferation of skills beyond the occupation(s) they are typically associated with;

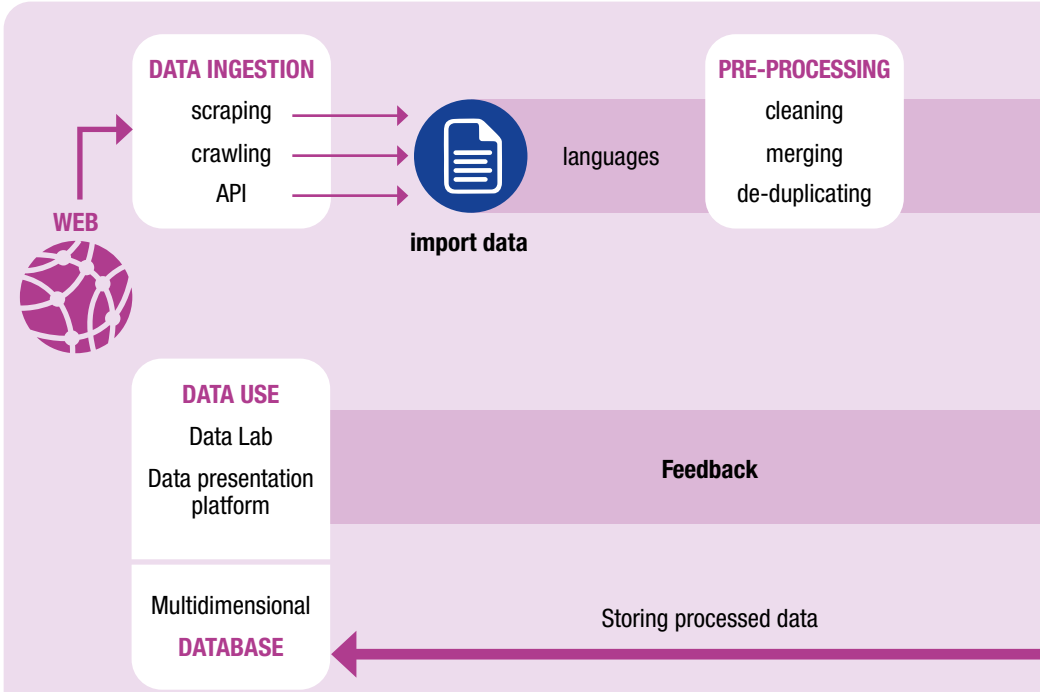
(e) synonyms: OJAs can give insight into new terms employers use to describe the same (set of) skills. This can help enrich existing skills taxonomies;

(f) job transitions: as OJAs can be used to map which skills and employment conditions are similar between different occupations, they can be used to shed light on potential job transitions within and across occupations and pay levels. CVs tend to list most or all jobs individuals have held in their careers. Such information can be used to understand school-to-work transitions, experience gains, career progression and occupational transitions. Analysis that links career moves and information on skills development after formal initial education or training can also inform career guidance.

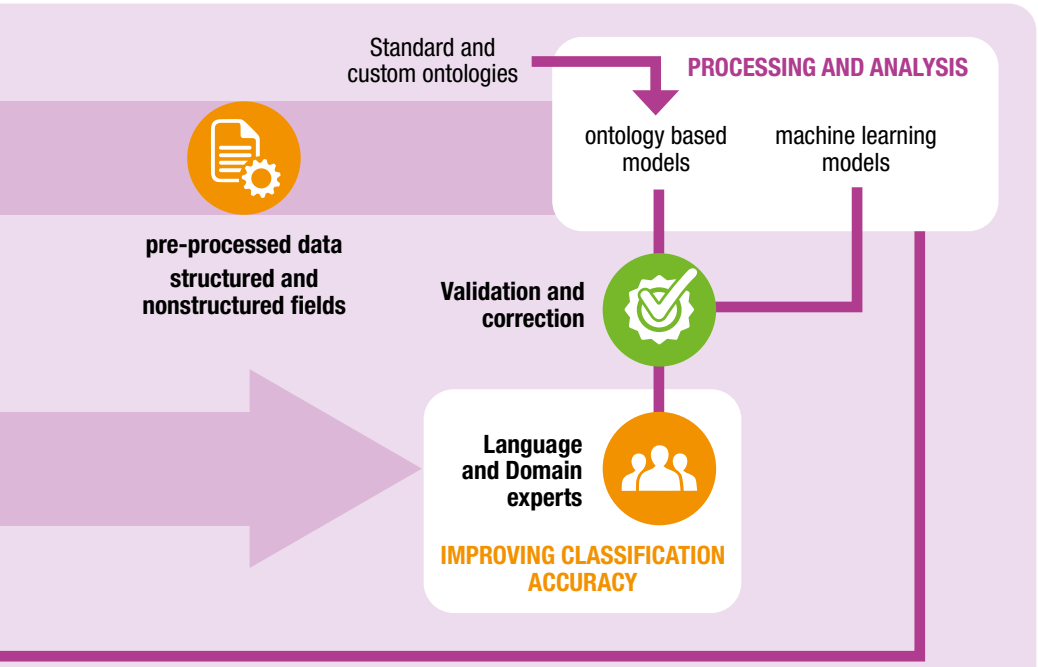
Figure 2. **Types of information typically contained in online job advertisements and CVs**

Source: IAG-TVET working group.

Box 3. **Collecting and preparing big data for analysis: data production systems**



Data ingestion is gathering primary documents from the web via web scraping (extracting structured data from websites), web crawling (systematically browsing web portals and downloading their pages) or an API (application programming interface) to download data directly from the database powering a web portal. While using an API is the preferred ingestion method as the data collected is of higher quality and can be downloaded much faster, it requires formal agreement with the website owner.



Data pre-processing is the process of making information of different quality and content suitable for analysis. It involves data cleaning (taking out irrelevant information or 'noise'), data merging (combining information on the same data point – e.g. a job or a person – from different sources) and data deduplication (removing information that appears in identical form in two or more web sources, such as a job advertisement appearing on several online job portals).

A DPS depends on standard ([8]) and custom ([9]) ontologies for processing and analysis of documents. Exact text matching, text similarity or machine-learning algorithms ([10]) can be used to allocate document content to skill, occupation, industry, region of the workplace, type of contract, and other categories. To remain relevant and accurate, ontologies should be continuously updated and enriched using automated techniques to reflect labour market and skills trends ([11]). Domain and language experts validate the machine-powered categorisation and propose corrections. Ontologies can also be updated manually to incorporate such trends, either for particular occupations or for an entire ontology such as ESCO.

Processed data are stored in a multidimensional database, which usually feeds a data presentation platform (DPP) to help users without big data expertise navigate the data using a graphical interface and a data entry point for more advanced users. A data lab provides experts with an easily and low cost-solution to use the information for basic or advanced data science analysis.

Source: IAG-TVET working group and Cedefop (2019b).

The 'bottom up' information contained in web-based big data is its main added value. The more detailed information on skills, occupations and careers, qualifications and other job requirements and characteristics in online job advertisements and CVs opens up many opportunities to strengthen labour market and skills intelligence (LMSI) (Table 1). Trends analysis can be undertaken because data can be collected frequently. Such work, however, requires sustaining a stable and consistent pool of online sources and overcoming continued operational complexities (such as regular monitoring of web scraping performance and updating of taxonomies). This requires resources to ensure continued tracing over time and can be challenging given that the online labour market is quite dynamic.

([8]) Standard ontologies refer to established classifications maintained by external organisations, such as ISCO for occupations, ESCO for skills, ISIC for industry, NUTS for geographical unit, ISCED for educational level.

([9]) Developed based on information available in documents, for example type of contract, experience, salary or working hours in OJA data, or type of course in education and training offers.

([10]) The machine learning algorithm uses statistical techniques to give computers the ability to 'learn' (progressively improve performance on a specific task) without being explicitly programmed.

([11]) ESCO is currently being updated with a view to releasing its next major version (version 1.1) by the end of 2021.

Table 1. **Potential of web-based big data for labour market actors**

	Governments (national, regional)	Education and training providers	Employers	Individuals
Information potential	More real-time monitoring of labour market trends More detailed and regionally adapted skills anticipation Additional insight into what broader trends mean for skills demand and supply	Faster insight into changes in professional practice More insight into demand for occupations and skills in the regional context Better understanding of regional skills supply and demand imbalances	More real-time insight into emerging trends in occupations and skills Additional insight into regional labour market situation Better understanding of critical skills bottlenecks	Better understanding of skills increasingly in demand More contextually adapted (region/sector) labour market and skills intelligence Increased understanding about labour market opportunities
Policy/action potential	Stronger feedback loops More proactive skills policy More responsive up-/reskilling measures More evidence-based competitiveness policy/strategy	Increased capacity to respond to changing skills needs More evidence-based programme offer More effective and proactive careers information and guidance provision	Increased capacity to map corporate skills needs More evidence-informed recruitment policy More forward-looking approach to training and development	More informed education and training choices More proactive approach to developing career and employability More successful transitions between occupations

Source: IAG-TVET working group.

CHAPTER 4.

Overcoming challenges and limitations

Unleashing the full potential of big data in developing better skills policies demands awareness of its unique challenges and limitations. These range from strategic and resource challenges, statistical and conceptual issues, to ethical and legal concerns (Table 2). Understanding these constraints is a precondition for developing a big data initiative, using it to produce sound evidence, correctly interpreting results and using them for policy-making purposes.

After a decision to pursue big data analysis, a fundamental choice needs to be made: buying data from a commercial vendor or developing a data collection and analysis system from scratch? The benefit of being able to act faster needs to be weighed against the disadvantages of relying on data which might lack transparency with regards to how they have been collected and processed. When it is possible to invest more time in system development, control over the entire production process is important, and costs linked to relying long-term on a vendor can outweigh the set-up and maintenance costs of a dedicated solution, developing a complete system appears to be the best choice ([12]).

The chosen approach has implications for the human resources needed. Analysis is impossible without well-trained big data experts and setting up a big data production system requires a wide variety of expertise. Developers design the infrastructure in-house or in the cloud. Experts with high-level programming skills are crucial for system set-up, regular monitoring, maintenance and further development. As recent experience with deploying big data analyses in education systems shows, just as important are domain experts. They are essential in developing ontologies and for training and correcting machine learning algorithms. As big data systems produce

([12]) An example of an own-generated, web-sourced, big data infrastructure is the European online vacancy analysis tool for Europe (Skills-OVATE), developed by Cedefop.

enormous amounts of data ([13]), a good data strategy, retention policy and controls safeguarding data security need to be developed.

Table 2. **Key challenges in developing and using web-based data for skills analysis**

Challenge	Core underlying issue	Main implications
Aligning aims and strategic choices	Deciding on make (building a collection and analysis system from scratch) or buy (external big data provider)	Full transparency, flexibility and oversight requires a customised system. Buying big data is an option when being fully in control is less important
Securing resources	Big data analysis requires specific technical and domain expertise and a dedicated hard- and software infrastructure	Big data analysis expertise, programming expertise for system set-up, monitoring, maintenance and development, domain expertise to develop ontologies and for algorithm training and supervision, system development expertise to advise on architecture (servers on premises, cloud solutions). Data strategy, retention policy and controls needed to safeguard security
Ethical and legal challenges	In gathering, storing, processing, analysing and presenting big data, ethical standards and privacy regulation must be respected	Website owners must be informed: it is advisable to reach agreement to avoid disruption of their activities. Arrangements must be in place to comply with applicable data protection legislation
Statistical biases	Unstructured and non-random nature of big data	Difficulties in applying standard statistical techniques and tests and in drawing generalised inferences about underlying populations
Representativeness	Varying coverage in labour market segments	Using big data to understand phenomena in particular labour market segments may be easier than developing representative statistics for them
Context-driven nature	Big data are shaped by the context in which it originates	Limited cross and within-country comparability and challenges in using multilingual taxonomies

Source: IAG-TVET working group.

([13]) For example, Cedefop's Skills OVATE tool is powered by about 15 TB data in various stages and formats.

Although information online is in the public non-regulated domain, ethical standards of use must be respected, as regularly extracting information may interfere with the functioning of job portals, CV tools, education and training provider websites, and other online sources. Agreements with website owners can help prevent this. Extracting skills information from CVs and other sources containing personal information is subject to data processing constraints. It must be ensured that all data documentation is up to date and national and international data protection legislation is respected.

Statistical biases, representativeness and conceptual challenges complicate big data analysis and make drawing valid conclusions challenging. Big data are unstructured and usually non-random and tend to cover particular labour market segments (such as highly skilled or ICT occupations in public portals or lower-skilled positions in some public employment service sources) better than others: Box 4 illustrates this for online job advertisements. Benchmarking the representativeness of big data against alternative representative data sources is, therefore, an important step in preparing analysis.

Box 4. **Representativeness challenges of big data based on online job advertisements**

- Not all jobs are advertised online and not all job advertisements lead to actual job openings. The nature and the maturity of the online labour market is shaped by the size of the informal economy, cultural factors and digital divides in internet connectivity and digital skills.
- Employers tend to use occupation-specific hiring strategies. High level professionals are often recruited via dedicated or privately owned portals or job-hunting. Public employment service (PES) portals are typically used for medium or low-skilled jobs.
- Some jobs are rarely advertised online at all, because word of mouth or a notice in a shop window are more effective and cheaper solutions for recruiting staff.
- Some portals restrict access to particular groups, such as the registered unemployed in the case of several national PES portals in the EU.
- Skills requested in the OJAs are not skills profiles. Employers emphasise the skills that give candidates a competitive edge and those that may help reduce the pool of available applicants. Lack of common standards and tools for describing skills in OJAs causes selectivity and variation in the skills indicated.

Source: Cedefop (2019a, 2019b).

Information gathered from CVs often has similar biases. Typically, only active jobseekers submit their CVs and keep them up to date. People in less ICT-intensive jobs are less likely to keep their online CVs updated. Such factors may result in sizable underrepresentation of particular groups. In developing countries with less mature online labour markets, such challenges typically play an even larger role (Box 5).

Box 5. **Challenges in using big data for skills analysis in developing countries**

The allure of a big data system in developed but mainly lower-income countries is that it may be easier to develop and maintain than traditional surveys. But in many developing countries, big data expand into data collection systems that are often less robust and weak in terms of capturing labour market developments, due to factors like low coverage (overall or sectoral) and high levels of informality. According to 2016 ILO estimates, the share of informal employment is 25.1% across Europe and Central Asia, compared to 85.8% across Africa or 68.2% across Asia and the Pacific (ILO, 2018). High informality will make it much less likely for jobs to be advertised online; as a result, the share of the labour market covered by OJV data sources will be low. Further, given that the share of informal jobs tends to be lower for higher-skilled positions, job vacancy data in developing countries will tend to be more biased towards these types of jobs (ILO, 2018).

Growing internet penetration and smartphone coverage in developing countries is opening doors to the development and wider use of labour market information based on big data. Despite the positive trends, low internet penetration and the limited use of online platforms by employers and workers in many developing countries remains a major barrier to sound analysis and reliable results. According to ITU estimates, only 28% of individuals across the African continent were using the internet, compared to 82.5% in Europe and 77.2% across the Americas.

Without anchors to surveys and administrative data, the risks of biases are high. Further challenges to using big data in developing countries typically include the identification of reliable online job portals, their coverage (and issues relating to lack of data, for example, on blue-collar jobs or informal employment), data-cleaning problems, taxonomy decisions and limited complementarity with other LMI (which is necessary to establish holistic skills governance). Organising support to data governance mechanisms, to capacity building and to development of backbone surveys to be deployed alongside big data is another key challenge for developing countries.

Source: IAG-TVET working group.

Big data are also strongly influenced by the context (labour market, education and training system, language and culture) in which it is generated. Where professions are heavily regulated, skills requirements are well understood because they are part of qualifications or regulated by professional organisations. As a result, OJAs are not as rich in terms of skills as in countries with more liberal systems ([14]). Digital divides and economic structure must be considered when comparing results between countries and regions. It is advisable to focus analysis on topics and areas that are well-represented by the data. It is easier to develop insight into digital technologies and their skills implications than to extract representative statistics on digital skills demand covering the entire labour market.

In summary, when analysing and presenting big data, especially to the public, researchers and policy-makers must always bear in mind the nature and character of such data to avoid misinterpretation. It is not advisable to use big data as a substitute for mainstream labour market analyses, because of the many challenges and limitations such data entail. While using web-based big data to complement labour market and skills intelligence (LMSI) is encouraged, it is important to act prudently when combining such data with evidence based on conventional statistical sources.

Sound understanding of big data is also important when combining different types of big data analysis. Attempts to combine information extracted from individual CVs with data on skill demands based on OJAs, aimed at developing 'mismatch' indicators, is a case in point. Combining information that has been developed based on biased samples, failure to acknowledge different motivations underlying such samples, and other ecological fallacies suggest caution in performing such matching exercises.

Incorporating new big data sources in national skills anticipation systems along traditional LMSI sources inevitably poses challenges for skills governance systems ([15]). Successfully integrating them into the policy-

([14]) For instance, in some highly regulated countries a vacancy saying 'plumber needed' does not need to elaborate much more in terms of the skills required; in others, without such regulations, a list of extensive skills requirements may need to be specified by employers to filter out appropriately trained candidates.

([15]) See for example:
ILO-Cedefop-ETF-OECD (2017). *Skill needs anticipation: systems and approaches*.
Cedefop's *Anticipating and matching skills* project that has assessed skills governance in four EU countries.
OECD (2016). *Getting skills right*.
OECD (2020d) *Strengthening the governance of skills systems*.

making process requires developing national/subnational big data strategies, incorporating stakeholder needs (including national statistical service, research institutions, web platforms, skills analysis units), interoperability between diverse data channels, and policies for data integration across different layers of governance. This requires substantial human and financial resources. To avoid difficulties in linking information sources, it is advisable and good practice to use existing and internationally accepted taxonomies and ontologies (ESCO, O*NET, ISCO, NACE and others).

CHAPTER 5.

Prospects for data-informed skills policy

The expectations for web-based big data as a game changer in skills analysis are high because of its potential in offering more detailed information in (quasi) real time. But there are positive and negative sides to the experience so far. Several tools are already assisting policy-makers and practitioners in labour market and learning settings and play a role in improving TVET delivery (Box 6).

Box 6. **Using big data to inform skills and TVET practices, policies and strategies**

In several countries, big data analysis is becoming an essential component of skills assessment and anticipation exercises that measure shortages in occupations or skills. For example, by combining information from online job advertisements with other traditional data sources and bottom-up quantitative or qualitative inputs, the UK's Migration Advisory Committee and the Malaysian Critical Occupations List assess which skilled occupations are in shortage.

One of the actions of the European Commission's European skills agenda is *Strengthening skills intelligence*. This includes building on Cedefop's Skills OVATE project to create a permanent online tool which includes real-time information for all interested stakeholders. The data Skills OVATE provides on emerging occupations and skills trends and the interrelationships between occupations and skills will also be used to support, scale up and automate ESCO maintenance and updates. The Skills agenda also envisions skills intelligence tailored to user needs on Europass, the EU platform supporting people to manage their learning and careers. It also seeks to promote the participation of social partners in generating skills intelligence and the increased use of skills intelligence by public employment services.

In Malawi and Myanmar, UNESCO mined data from local online job portals and job titles were mapped using JobKred's taxonomy. The frequency of mapped job titles was used as an indicator of employer demand in the labour market. Subsequently,

the job descriptions were processed by JobKred's predictive engine to identify the relevant skills ([16]). UNESCO plans to use big data from online job portals to support evidence-based TVET policies in the *Youth employment in the Mediterranean* (YEM) project. A partnership with McKinsey on testing the power of advanced analytics (AA) in education is in place. UNESCO recommends that governments and stakeholders ensure ethical, transparent and auditable use of data, be cognisant of the dilemmas of balancing open access to data and data privacy protection. It advocates being mindful of the legal issues and ethical risks related to data ownership, data privacy and public data availability (UNESCO, 2019c).

In ETF partner countries the use of digital tools and online portals – public and private – for posting and managing job vacancies is growing. In parallel, big data are used to detect emerging skills needs in economic sectors. The ETF has launched a discussion with national stakeholders about possible actions that might help developing and transition countries in tapping the potential of big data for labour market intelligence. Such actions include: national feasibility studies to identify, validate, and rank web sources; developing real-time labour market information collection systems; and defining data analytics models to support decision-makers in policy design and evaluation. Publicly accessible dashboards presenting 2020 information for Tunisia and Ukraine were released in November 2020. Upon positive assessment of this first full-scale big data pilot in two countries, the project will continue in 2021 with further data collection and analysis and inclusion of a third country. Patent and bibliometric analysis been successfully applied in Israel (agri-tech), Turkey (automotive) and Morocco (agri-business) to understand better the skills needs associated with emerging technologies ([17]).The use of big data in these countries has produced highly relevant findings, which have been validated and complemented by national stakeholders.

Skills assessment and matching tools increasingly rely on big data. The Flemish public employment service uses text analytics on OJAs and CVs to match open positions to jobseekers and to identify training opportunities that bridge the gap between jobseeker skills and what an employer demands. The Competent database, which was developed with social partners, is used as a backbone to link qualifications and work experience to skills requirements. The approach has also been successfully implemented in Malta. The Amsterdam House of Skills – a public-private partnership aiming to facilitate mobility to and between work in the region – uses the ESCO and O*Net taxonomies to set up big-data-powered skills development and matching tools to address skills bottlenecks and support transitions.

([16]) For more details, see UNESCO (2019a) Malawi TVET Policy review and UNESCO (2019b) Myanmar TVET System review.

([17]) The approach will be implemented in six other countries in 2021 (energy sector in Albania and Tunisia, health and care sector in Ukraine, other cases to be decided).

Career guidance providers and online career platforms are making more use of big data to provide relevant and timely labour market information that can help students and adults make informed education and career choices. In New Zealand, the Ministry of Business, Innovation and Employment provides information based on the trend in the number of online job advertisements per occupation in Occupation outlook, its online education and career exploration tool.

Source: IAG-TVET working group.

On the positive side, web-based big data can provide valuable and generally quicker insights at a finer level of granularity in ways that are not always feasible when relying on conventional LMSI. With the digitalisation of societies, web-based big data are expected to become an integral part of the data infrastructure that supports countries, regions, employers, learners and education and training providers in understanding socioeconomic phenomena, including skills demand and supply trends in labour markets.

It is often claimed that big data systems can amortise their high set-up costs over time. While this is true to some degree, it is important to be aware of the continued need to allocate financial and non-financial resources required to maintain high-quality and up-to-date big data infrastructure. Developing a solid big data production system is a formidable technical challenge. Artificial intelligence – the basis for analytical tools – allows processing of enormous amounts of information but cannot function without proper training by domain experts. The need for domain-specific knowledge and expertise is often underestimated. Big data analysis is only as good as the underlying ontologies and building and maintaining them requires continuous effort.

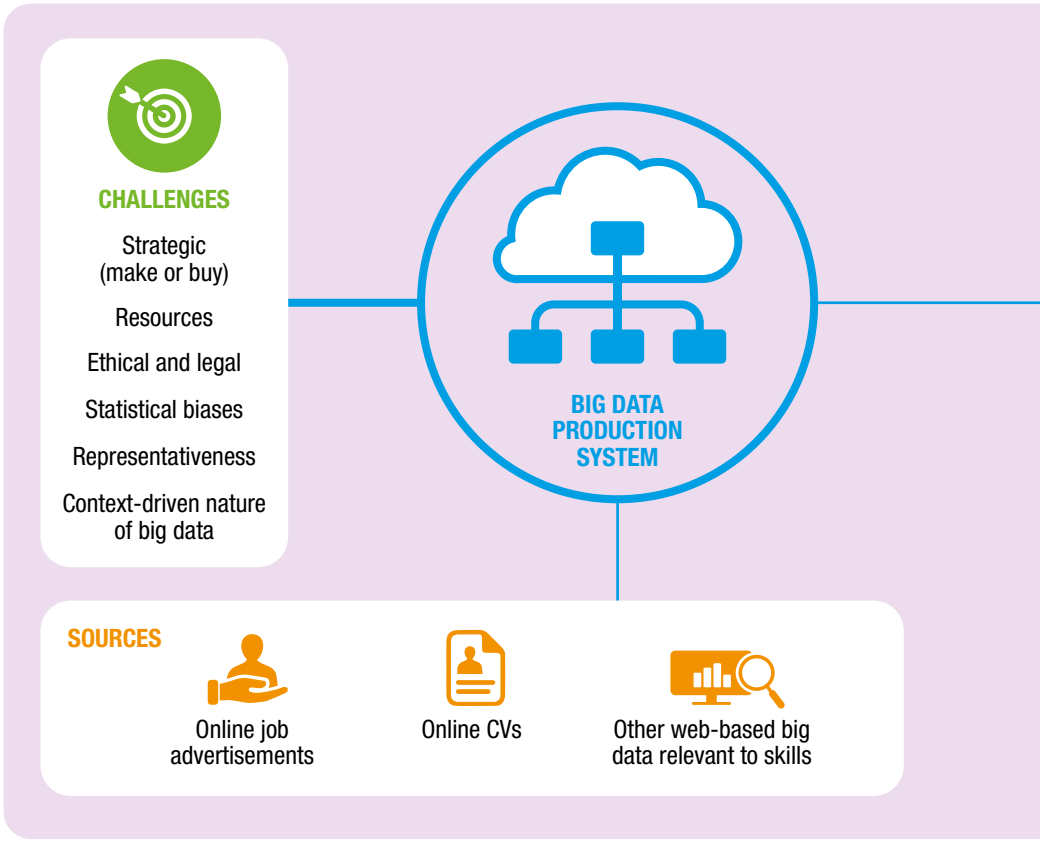
Although big data have clear and proven potential in the changing world of work, experience so far clearly demonstrates that they are not a panacea. Policy-makers must be cautious when interpreting the findings of big data analysis and using them to shape or adapt measures or strategies. It is also important to ensure investing in big data analysis does not come at the expense of LMSI based on forecasts, surveys, administrative data or skill foresights, the more so because the potential of big data only fully materialises when combined with such LMSI.

Progress in gradually integrating big data in conventional LMSI systems depends on the capacity to discover areas of application in national and regional contexts. Statistical departments, research centres and analytical

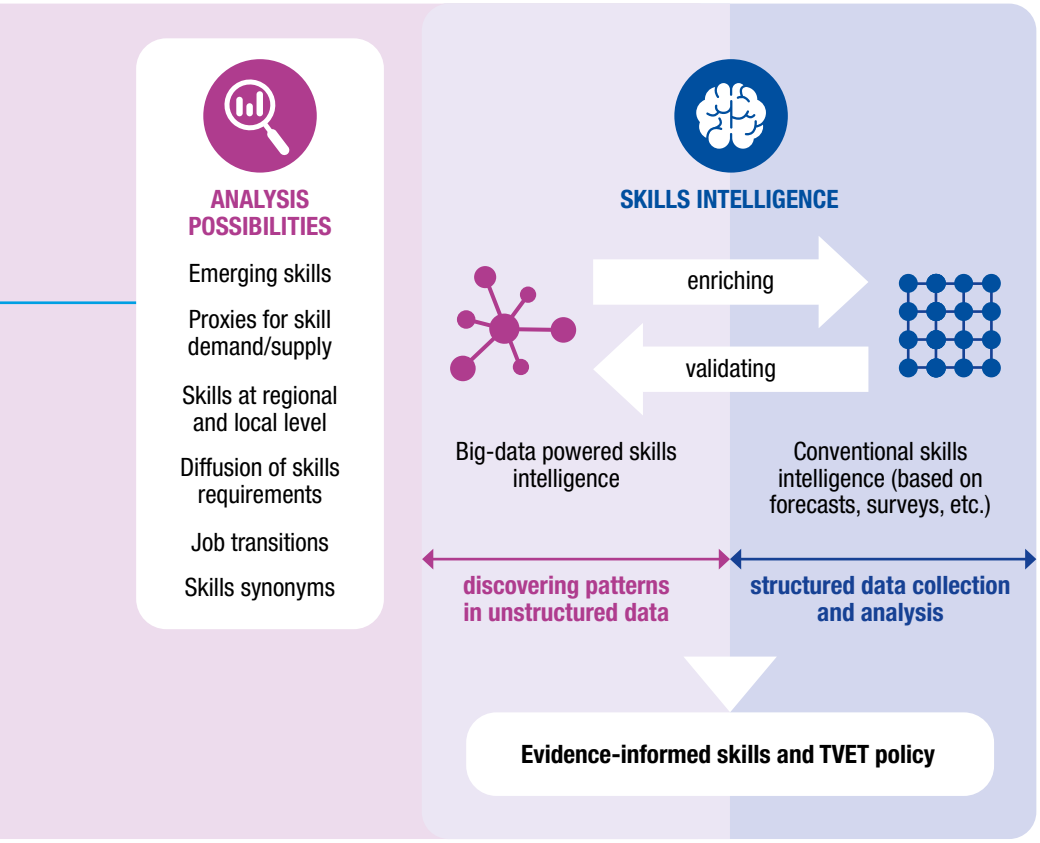
units in government should be well informed of the possibilities and value added of big data, as well as the challenges. The data science methods and AI powering big data systems should be accessible by the research community and government entities and open to scrutiny. Analysis and dissemination of good practice in the application of big data methods in LMSI can be organised as inter-institutional projects. In such settings, countries, regional communities, and (international) research bodies and organisations can jointly work towards promoting their benefits and wider use.

Big data are the engine of artificial intelligence, but 'human intelligence' is and will remain essential. Skills-relevant information online must be classified and checked by machines but the human in the loop trains and improves the machine-learning algorithms. Big data analysis, despite its AI glamour, is resource-intensive and is not possible without human expertise: in collection, analysis, validation and interpretation. Big-data informed LMSI cannot rely on computational and algorithmic power alone. It is the combination of artificial and human intelligence that will be central to developing big data's role in shaping effective TVET and skills policies in the coming years.

Figure 3. **In short: combining big data-powered and conventional skills intelligence**



Source: IAG-TVET working group.


ANALYSIS POSSIBILITIES
Emerging skills
Proxies for skill demand/supply
Skills at regional and local level
Diffusion of skills requirements
Job transitions
Skills synonyms
SKILLS INTELLIGENCE
enriching
validating
Big-data powered skills intelligence
Conventional skills intelligence (based on forecasts, surveys, etc.)
discovering patterns in unstructured data
structured data collection and analysis
Evidence-informed skills and TVET policy

Acronyms

AI	artificial intelligence
API	application programming interface
DPP	data presentation platform
DPS	data production system
ESCO	European taxonomy for skills, competences and occupations
IAG-TVET working group	Inter-agency technical and vocational education and training (IAG-TVET) working group on skill mismatch in digitised labour markets
LMI	labour market intelligence
LMSI	labour market and skills intelligence
NLP	natural language processing
OJA	online job advertisements
OVATE	online vacancy analysis tool for Europe
PES	public employment service
TVET	technical and vocational education and training

References

Cedefop (2019a). *The online job vacancy market in the EU: driving forces and emerging trends*. Luxembourg: Publications Office. Cedefop research paper; No 72. http://data.europa.eu/doi/10.2801/16675

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

ETF (2019). *Big Data for labour market intelligence: an introductory guide*. Turin: ETF. https://www.etf.europa.eu/sites/default/files/2019-06/Big%20data%20for%20LMI.pdf

ILO (2018). *Women and men in the informal economy: a statistical picture* (3rd ed). Geneva: ILO. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_626831.pdf

ILO (2020). *The feasibility of using big data in anticipating and matching skill needs*. Geneva: ILO. https://www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_ent/documents/publication/wcms_759330.pdf

Laney, D. (2001). *3d Data management controlling data volume velocity and variety.* https://idoc.pub/documents/3d-data-management-controlling-data-volume-velocity-and-variety-546g5mg3ywn8

Bibliography and further reading

ILO (2019). *Skills for a greener future: a global view based on 32 country studies.* Geneva: ILO. https://www.ilo.org/skills/pubs/WCMS_732214/lang--en/index.htm

ILO; Cedefop; ETF; OECD (2017). *Skill needs anticipation: systems and approaches*. Geneva: ILO. https://unesdoc.unesco.org/ark:/48223/pf0000371392?posInSet=1&queryId=aa2a1748-3f35-4995-aa6c-404ca1dda263

OECD (2016). *Getting skills right: assessing and anticipating changing skill needs, getting skills right*. Paris: OECD Publishing. https://doi.org/10.1787/9789264252073-en

OECD (2017). *Digital economy outlook 2017.* Paris: OECD Publishing. https://doi.org/10.1787/9789264276284-en

OECD (2019a). *Benchmarking higher education system performance.* Paris: OECD Publishing. Higher education series. https://doi.org/10.1787/be5514d7-en

OECD (2019b). *Measuring the digital transformation: a roadmap for the future.* Paris: OECD Publishing. https://doi.org/10.1787/9789264311992-en

OECD (2020a). *Employment outlook 2020: worker security and the COVID-19 Crisis.* Paris: OECD Publishing. https://doi.org/10.1787/1686c758-en

OECD (2020b). *Skills measures to mobilise the workforce during the COVID-19 Crisis – OECD policy responses to coronavirus*. https://read.oecd-ilibrary.org/view/?ref=135_135193-hgf8w9g731&title=Skill-measures-to-mobilise-the-workforce-during-the-COVID-19-crisis

OECD (2020c). *Labour market relevance and outcomes of higher education in four US States: Ohio, Texas, Virginia and Washington.* Paris: OECD Publishing. Higher education series. https://doi.org/10.1787/38361454-en

OECD (2020d). *Strengthening the governance of skills systems: lessons from six OECD countries.* Paris: OECD Publishing. OECD skills studies. https://doi.org/10.1787/3a4bb6ea-en.

OECD (forthcoming). *Measuring the impact of the COVID-19 crisis on jobs and skills demand.* OECD Policy responses to coronavirus.

OECD (forthcoming). *OECD skills outlook 2021*.

UNESCO (2019a). *Malawi TVET policy review.* https://unesdoc.unesco.org/ark:/48223/pf0000367974?posInSet=1&queryId=1a2f1e59-2648-4a4f-a383-6545a93ff8d1

UNESCO (2019b). *Myanmar TVET system review* https://unesdoc.unesco.org/ark:/48223/pf0000371392?posInSet=1&queryId=aa2a1748-3f35-4995-aa6c-404ca1dda263

UNESCO (2019c). Beijing Consensus on artificial intelligence and education: outcome document of the *International Conference on Artificial Intelligence and Education 'Planning education in the AI era: lead the leap', Beijing, People's Republic of China, 16 to 18 May 2019*. https://unesdoc.unesco.org/ark:/48223/pf0000368303

Web pages and online tools

Cedefop – Anticipating and matching skills https://www.cedefop.europa.eu/en/events-and-projects/projects/assisting-eu-countries-skills-matching

Cedefop – European online vacancy analysis tool for Europe (Skills-OVATE) https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies

Cedefop – News page presenting the first comprehensive compendium of guides on skills anticipation methods, produced by ETF, Cedefop and ILO. https://www.cedefop.europa.eu/en/news-and-press/news/first-comprehensive-compendium-guides-skills-anticipation-methods

European Commission – ESCO https://ec.europa.eu/esco/portal/home

OECD – Science, technology and industry scoreboard: the digital transformation. https://doi.org/10.1787/9789264268821-en

2229 EN – TI-09-21-027-EN-N - doi:10.2801/25160

CEDEFOP | European Centre for the Development of Vocational Training

Europe 123, Thessaloniki (Pylea), GREECE
Postal address: Cedefop service post, 570 01 Thermi, GREECE
Tel. +30 2310490111, Fax +30 2310490020, Email: info@cedefop.europa.eu

visit our portal **www.cedefop.europa.eu**

Publications Office of the European Union

ISBN 978-92-896-3235-5