# The feasibility of using big data in anticipating and matching skills needs

# ▶ The feasibility of using big data in anticipating and matching skills needs

ILO Geneva

# ▶ Editors and authors

## Editors

**Ana Podjanin**
International Labour Office (ILO)

**Olga Strietska-Ilina**
International Labour Office (ILO)

## Authors

Introduction: **Ana Podjanin**, ILO

Section 1.1: **Konstantinos Pouliakas** and **Jasper Van Loo**, Cedefop

Section 1.2: **Eduarda Castel-Branco**, European Training Foundation (ETF)

Section 1.3: **Fabio Mercorio**, University of Milan – Bicocca

Section 1.4: **Stefan Winzenried**, Janzz Technology

Section 2.1: **Claudia Plaimauer**, 3s

Section 2.2: **Renier Van Gelooven**, SBB – Foundation for Cooperation on Vocational Education, Training and the Labour Market

Section 2.3: **Gábor Kismihók**, Leibniz Information Centre for Science and Technology University Library (TIB)

Section 2.4: **Labour Market Information Council, Canada, Employment and Social Development Canada and Statistics Canada** (presented by Tony Bonen, LMIC)

Section 2.5: **Andy Durman**, EMSI UK

Section 3.1: **Carlos Ospino Hernández**, Inter-American Development Bank

Section 3.2: **Sukriti**, LinkedIn India

Section 3.3: **Hiromichi Katayama**, UNESCO

Section 4.1: **Inna Grinis**

Section 4.2: **Julia Nania**, **Hal Bonella**, **Dan Restuccia** and **Bledi Taska**, Burning Glass Technologies

Section 4.3: **Ana Podjanin**, **Olga Strietska-Ilina** and **Bolormaa Tumurchudur-Klok**, ILO

Conclusions: **Cornelius Gregg**, **Olga Strietska-Ilina**, ILO

# ▶ Foreword

Mismatches between the skills offered and those required on the job market continue to be high on the policy agenda of both developed and developing countries across the world. Digitalization and technological disruptions are changing skills demand very fast, turning the task of identifying skills needs into the pursuit of a fast-moving target that is hard, if not impossible, to hit. At the same time, uncertainty and disruption raise the bar of expectation in predicting the skills required by the jobs of the future even higher. The current publication is based on a discussion among experts about how big data analytics might be used to help anticipate skills needs better and faster. This discussion took place at an ILO workshop in Geneva, Switzerland, on 19–20 September 2019, just a couple of months before the beginning of the COVID-19 outbreak. At the time of publication, we are already seeing unprecedented disruption in the labour market, along with unprecedented levels of public expectation for rapid answers to complex questions: what skills are needed, what reskilling measures deserve budgetary allocations, which active labour market measures should be prioritized and how to advise those who have lost their jobs about possible career prospects.

The traditional methods of skills needs anticipation and matching involve reliance on either quantitative analysis or qualitative research. Quantitative approaches typically use proxies for the measurement of skills, such as occupations, qualifications and levels or types of education. Such proxies provide useful information but are not sufficiently informative about the specific skills and competencies needed on the labour market. Without this extra level of information, skills remain hard to pin down in policy-making. Qualitative approaches certainly fit the purpose better, allowing us to identify specific skills and competency needs at regional or sectoral level, or for specific occupations and qualifications. However, they are fairly time-consuming and require significant resources; also, given the speed at which labour markets are changing, they run the risk of producing information that is obsolete before it can be used. This is why researchers and policy-makers are looking for other sources of information that will help to address the problem more efficiently.

The increasing use of the Internet for publishing job vacancies offers an incredibly rich source of data. Namely, it allows us to access in real time information on current skills demand, captured through job descriptions. Since the information is already there, its use is also efficient in terms of cost. However, the data from this source lacks structure, suffers from duplications and lack of representativeness, needs cleaning and quality checking, and is subject to many other potential problems, including data privacy issues that stand in the way of its effective use. In developing countries, an additional limitation is a limited reach of online vacancies due to poor connectivity and a large share of informal jobs. Nevertheless, online job vacancies and other types of big data analytics have great potential to contribute to a better understanding of labour markets, especially if complemented by more traditional sources of information.

This publication is composed of the contributions to the ILO workshop on the use of big data for skills anticipation and matching. The aim of this workshop was to share good practices and experiences, and identify to what extent these existing methods and approaches can be used and adapted for developing countries.

**Srinivas Reddy**
Chief, Skills and Employability Branch,
ILO Employment Policy Department

# ▶ Acknowledgements

# ▶ Contents

**4. Connecting the dots: Combining big data and other types of data to meet specific analytical demands**                                                                                   **77**

**Conclusion**                                                                                                       **94**

**References**                                                                                                      **96**

**List of figures**

## List of tables

## List of boxes

# ▶ List of abbreviations

| | |
|---|---|
| **AI** | artificial intelligence |
| **API** | application programming interface |
| **Cedefop** | European Centre for the Development of Vocational Training |
| **DSR** | digital skills rate |
| **e-CF** | European e-Competence Framework |
| **ESCO** | European skills/competencies, qualifications and occupations framework |
| **ESDC** | Employment and Social Development Canada |
| **ETF** | European Training Foundation |
| **ETL** | extract, transform, load |
| **EU** | European Union |
| **HR** | human resources |
| **ICT** | information and communications technology/ies |
| **IADB** | Inter-American Development Bank |
| **ISCED** | International Standard Classification of Education |
| **ISCO** | International Standard Classification of Occupations |
| **ISIC** | International Standard Industrial Classification |
| **JMI** | Job Market Intelligence (Cedefop) |
| **KDD** | knowledge discovery in databases |
| **LA** | learning analytics |
| **LMI** | labour market information |
| **LMIC** | Labour Market Information Council (Canada) |
| **NACE** | Nomenclature statistique des activités économiques dans la Communauté européenne (statistical classification of economic activities in the EU) |
| **NLP** | natural language processing |
| **NOC** | National Occupational Classification (Canada) |
| **NSO** | national statistical office |
| **NUTS** | Nomenclature of Territorial Units for Statistics |
| **OECD** | Organisation for Economic Co-operation and Development |
| **OVATE** | Online Vacancy Analysis Tool for Europe |
| **OJVs** | online job vacancies |
| **p.a.** | per annum |
| **PES** | public employment service(s) |
| **SBB** | Foundation for Cooperation on Vocational Education, Training and the Labour Market (Netherlands) |

| **SOC** | Standard Occupational Classification (US/UK) |
| --- | --- |
| **STC** | Statistics Canada |
| **STEM** | science, technology, engineering and mathematics |
| **TVET** | technical and vocational education and training |
| **UNESCO** | United Nations Educational, Scientific and Cultural Organization |
| **VET** | vocational education and training |
| **WEF** | World Economic Forum |

# ▶ Introduction

In dynamic and constantly changing labour markets, identifying skills needs is a significant challenge. Imbalances in the labour market, reflected in difficulties businesses face in sourcing the skills they need, a high incidence of skills mismatches, and significant unemployment or underemployment, especially among youth, are observed in most countries, albeit in different forms and to different extents. In view of the rapidly evolving labour market, there is a need to address not only currently observed mismatches, but also those that could potentially appear in the future, if the labour force is not adequately prepared to meet future needs. In order to tackle these issues, policy-makers, employers, workers, providers of education and training, and students all need timely and accurate information about demand for skills on the labour market.

Traditionally, policy-makers have used information both from official labour force surveys and from other surveys to provide quantitative information on labour market needs. While these sources are very rich in information and can be nationally representative, they can also have significant limitations. In less advanced economies they may not be conducted on a regular basis because of their high cost. More generally, the indicators they provide, such as occupation or qualifications, are only proxies for understanding actual skills requirements, and may not by themselves convey enough specific information to reliably guide action. The skills needs associated with particular occupations vary with context, and change over time.

Emerging new sources of data on skills have the potential to provide real-time and detailed information on skills needs in a cost-effective way. Technological advances, digitalization and Internet platforms have made it possible to collect very large, and rich, data sets, or so-called "big data", for many purposes. Data on the content of job advertisements has been collected systematically from online job postings in a range of countries, contributing to the generation of huge data sets containing detailed information on the requirements advertised. Information typically recorded includes the specific skills needs stipulated and skills-related indicators included in advertisements, such as job titles, along with requirements for qualifications, certifications and experience, as well as other information about each vacancy such as the employer, the economic sector, the occupational category and the geographic location of the post advertised. Data derived from online job postings can be collected in real time and, in contrast with data from surveys that require time for processing before publication, can be used almost immediately. This immediacy also offers an advantage over skills requirements taxonomies, which usually take a considerable amount of research and analysis, along with time to be produced and regularly updated.

The richness of information featured in online job vacancies data sets has attracted considerable attention, and has underpinned many publications within both academia and international organizations.

However, it is important to take into account the limitations associated with using vacancy data derived from the Internet as a basis for labour market information (LMI). The sample of vacancies collected may not be representative of all online-advertised vacancies, and is unlikely to be representative of all job vacancies because of differences in recruitment practices by occupation and industry. Higher-level skills are more likely to be advertised online, especially in less advanced economies. Online job advertisements chiefly cover the formal economy, so skills needed for informal employment are likely to be under-represented.

Moreover, while skills requirements noted in online job listings provide a window on the landscape of detailed skills requirements, they do not provide full listings of skills required in the manner of an occupational competency standard. There are no consistent standards stipulating what skills should be included in an advertisement. Many skills may be omitted because they are taken for granted by the recruiter, while those included may have been identified by the employer as salient because they differentiate the job on offer from similar jobs. Other potential drawbacks inherent in real-time data arise from a lack of structure, imperfect information and measurement errors, especially related to duplicate observations, and advertisements in which the number of jobs available is unspecified. Legal and regulatory matters may also be an issue, especially in relation to data privacy, which also poses challenges to extending the use of big data in labour market analysis.

Analyses of vacancy data do not constitute a direct substitute for the main existing sources of labour market data. They provide flow-related measures inclusive of churn, whereas conventional labour market analysis is often based on measures of stock or on direct measures of flow that seek to filter out churn. It remains to be seen to what extent vacancy-based big data analysis will be mainstreamed into national skills anticipation systems, to what extent it will be used as a complement to or a substitute for other parts of these systems, and to what extent it could provide a fast track into skills anticipation for less developed countries that do not have established systems. The future role of big data in skills anticipation beyond the vacancies data domain also remains unclear.

This publication collects together the contributions presented during the ILO workshop "Can we use big data for skills anticipation and matching?", which took place on 19–20 September 2019 at ILO headquarters in Geneva, Switzerland. Discussions during the workshop considered the feasibility of using big data in the context of skills anticipation and matching, and both the potential and the limitations of big data in skills analysis. Participants had the opportunity to offer suggestions on how to advance the agenda in this area, to share good practices and to suggest solutions to commonly found challenges. While these methods of data analysis have already been extensively used in advanced economies, a particularly important focus of the discussions was how they can best be deployed in developing economies.

The first chapter of this publication presents the contributions related to the more conceptual and technical aspects of using real-time data in skills analysis. Following these reflections, the second chapter puts together best practices and experiences from advanced economies. Reflecting the importance of discussing the potential of the use of online job vacancy data in the context of developing economies, the third chapter offers insights from analyses carried out across Latin America and Asia. The fourth chapter shifts the focus away from specific country contexts towards new approaches, such as combinations of big data with other sources. It also opens the way for further discussion in relation to the interpretation of this type of data.

# 1
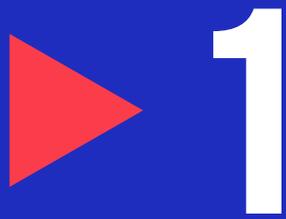
# Conceptual
# and technical aspects
# of knowledge-sharing
# on the usage of big data

This chapter discusses conceptual and technical aspects of big data analytics, and the related challenges and opportunities for improving LMI and delivering real-time and detailed skills demand analysis. It looks into the potential of the use of online job vacancies for skills needs anticipation and matching.
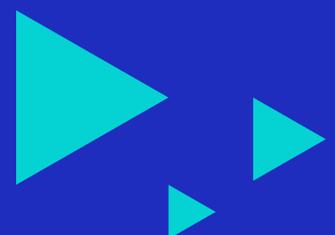
# ▶ 1.1. Cedefop and the analysis of European online job vacancies

**Jasper Van Loo and Konstantinos Pouliakas**

## 1.1.1. Introduction

EU decision-makers responsible for education and training, labour markets and/or skills need timely and reliable skills intelligence to support them in developing policies to better match skills with labour market needs. In the light of rapidly changing labour market needs, skills intelligence is crucial to designing, reforming and "future-proofing" education and training programmes. Surveys of employers, workers, graduates or the wider population can be, and have been, used to collect representative information on skills. But they are typically costly and time-consuming to implement, requiring substantial conceptual development in advance, and high response rates if they are to yield representative findings. Other "traditional" methods such as occupational and skills forecasts provide useful insights into medium- and long-term labour market trends; but, owing to the use of proxies for skills demand, endogeneity issues and time-lags between data collection and the generation of results, they are less suitable for prompt detection of employers' changing skills needs.

Since 2015, in line with its mandate to analyse labour market and skills trends in EU Member States, the European Centre for the Development of Vocational Training (Cedefop) has been investigating how information on skills demand available in online job vacancies (OJVs) can be used to generate faster and more detailed skills intelligence for the EU – as a complement to its other skills intelligence tools, namely the European skills forecast, the European skills and jobs survey and the European skills index.[1] A feasibility pilot study involving a limited number of countries, carried out in 2015, highlighted the potential for a pan-European online vacancy collection and analysis system to provide a detailed and unique set of policy-relevant information. In the past few years, Cedefop has focused on setting up the Skills OVATE (Online Vacancy Analysis Tool for Europe), a fully fledged system to collect and present indicators extracted from OJVs. Here we outline the Cedefop approach and provide an overview of the experience so far.

## 1.1.2. Analysing OJVs: Opportunities and challenges

OJVs are a rich source of detailed information about skills and other job requirements that are difficult to gather via traditional methods. Access to this information can help labour market actors better understand skills demand and its dynamics; enable individuals to make better career and skills development choices; support employers in developing or adjusting human resources (HR) policies; help policy-makers make more informed decisions; and improve the targeting of employment services, guidance counsellors and learning providers.

OJV analysis can provide additional, detailed and timely insights on labour market trends, and enables new and emerging jobs and skills to be identified early. But OJV analysis does not replace other types of labour market information and intelligence – it is in fact most powerful when combined with conventional sources. And it is crucial to acknowledge that meaningful analysis requires sound understanding of the online labour market in different countries and awareness of the key challenges in using OJVs for skills and labour market analysis. These challenges include:

▶ **representativeness:** vacancies in some sectors and occupations are over-represented in OJV portals, while those in other sectors are under-represented;

---

1 See, respectively: https://www.cedefop.europa.eu/en/publications-and-resources/data-visualisations/skills-forecast; https://www.cedefop.europa.eu/en/events-and-projects/projects/european-skills-and-jobs-esj-survey; https://skillspanorama.cedefop.europa.eu/en/indicators/european-skills-index [26 July 2020].

▶ **completeness:** skills listed in a vacancy notice do not reflect the full job profile – employers tend to list only critical skills and qualifications to "attract and filter" job applicants;

▶ **maturity:** patterns of use of OJV portals differ both across and within countries, reflecting the state of the "digital divide", skills shortages and particular employment structures;

▶ **simplification:** to be usable in skills and labour market analysis, vacancy notices must be machine-readable and use a standardized vocabulary – and, given the quantity of data, simplifying assumptions have to be made;

▶ **duplication and status of vacancies:** the same vacancy notice may be published on several websites and may not necessarily correspond to an actual job opening.

With the advent of greater digitalization, individual jobseekers are increasingly turning to online sources in looking for jobs. At the same time, technological change and skills shortages are leading employers to rely more often on web-based channels (e.g. online platforms) as a means of attracting key professionals or employees who possess specific skills and characteristics. In response to such drivers, the online job-searching market is expanding and very dynamic. Moreover, the dividing lines between online job portals and social media are getting more porous, boosting online labour markets even further. Nevertheless, the development and use of online job markets vary greatly across economic sectors, occupations and companies. For instance, in some countries employers are more likely to advertise job vacancies highlighting specific job titles and formal qualification criteria, whereas in others, posts advertising vacancies are more heavily weighted towards job-specific and transversal skills.

The advantages and disadvantages of using OJVs to analyse skills needs are summarized in table 1.1.

▶ **Table 1.1  Advantages and disadvantages of using OJVs for analysis of skills needs**

| Advantages | Disadvantages |
|---|---|
| ▶ (Almost) instant | ▶ Information is unstructured and imperfect |
| ▶ Big volume of information | ▶ Issues of non-representativeness |
| ▶ Time and cost effectiveness | ▶ Measurement errors (e.g. duplication or extended lifespan of vacancies) |
| ▶ In-depth information on skills and skill needs across and within units (e.g. countries, occupations) | ▶ Privacy concerns, ethical/legal considerations |
| ▶ No need to collect "new" data | ▶ Need for advanced data analytical skills (e.g. software, programming, coding) |
| ▶ Online information declared by individuals may be more "truthful" | ▶ Analysis confounded by selected taxonomy or clustering method(s) |
| | ▶ "Partial" occupational skills profiles |

**Source:** Authors.

## 1.1.3.  The online labour market in the EU

To understand better how OJVs are used across countries, Cedefop mapped and analysed the OJV landscape in all 28 EU Member States.[2] The share of vacancies published online in the EU ranges from under 50 per cent in Denmark, Greece and Romania to almost 100 per cent in Estonia, Finland and Sweden (see figure 1.1); there are also differences between regions within countries. With some countries reporting annual growth rates in online job markets of over 10 per cent, coverage is likely to increase in the near future.

Countries also differ in terms of the structure of their OJV market. In Denmark, Finland and Malta, for instance, the market is dominated by a few leading portals and public services tend to be influential players. Greece, Ireland, Italy and the United Kingdom, on the other hand, have more job portals with similar market power and public employment services are less influential compared to private actors.

---

2   This included the United Kingdom, which left the EU on 31 Jan. 2020. For more information, see Cedefop, 2019a.

▶ **Figure 1.1. Proportions of job vacancies published online in EU countries: Assessment by country experts, 2017 (%)**



**Note:** AT = Austria; BE = Belgium; BG = Bulgaria; CY = Cyprus; CZ = Czech Republic; DE = Germany; DK = Denmark; EE = Estonia; EL = Greece; ES = Spain; FI = Finland; FR = France; HR = Croatia; HU = Hungary; IE = Ireland; IT = Italy; LT = Lithuania; LU = Luxembourg; LV = Latvia; MT = Malta; NL = Netherlands; PO = Poland; PT = Portugal; RO = Romania; SE = Sweden; SK = Slovakia; UK = United Kingdom. Data for Lithuania and Slovenia not available. In Slovenia expert assessment states that online job announcements is a "prevailing" method.

**Source:** Authors.

Whereas formerly OJV platforms used to target predominantly highly skilled workers, today many advertise jobs for almost all occupations and skill levels. Two major trends are driving the use of OJVs:

▶ **Digitalization and rising computer literacy** have boosted both advertising of, and searching for, jobs online. Public employment services and private job portals increasingly use digital matching tools to make online recruitment more attractive to both jobseekers and employers.

▶ Economic growth has led to **skills shortages** in certain sectors and regions, driving employers to move their recruitment efforts beyond their traditional job applicant pools to reach potential candidates in other regions or countries.

While the online vacancy market is growing, not all types of job openings are published online. For example, specialists and executives may be headhunted; for some jobs (e.g. waiters) the employer may simply display a note in the workplace window, while for others employers prefer internal recruitment or word of mouth. To recruit scarce talent, employers approach young people in schools and universities directly and/or use various school-to-work transitional mechanisms (e.g. apprenticeships) to screen potential workers. Vacancies are likely to be advertised online more by large, international firms and those in manufacturing and services sectors, such as finance or information and communications technologies (ICT), and less by small firms and those in construction, agriculture and hospitality. Location also makes a difference. More online vacancies are registered in urban areas, where the concentration of employers in service sectors is higher, as is supply of skilled talent, and where information asymmetries are felt more acutely, compared to rural areas, where print media and word of mouth will probably continue to play a significant role.

## 1.1.4. Collecting and analysing online job vacancies

The process of collecting OJV data can be broken down into several steps (see figure 1.2).[3] The first is selection of the information source (on this step, see section 1.2.4 below). The identified sources then enter the

---

3  For a more extensive description, see Cedefop, 2019b.

"data ingestion" phase, which involves identifying vacancies available in the sources and downloading their content. During the next "pre-processing" phase, the downloaded data are cleaned of irrelevant content. A process of "information extraction" then translates the relevant content of the OJV into a database organized by taxonomies and classifications, which is used to present processed results.

▶ **Figure 1.2. A summary of the OJV data collection and production process**

| Selection | Ingestion | Processing | | Data use | |
|---|---|---|---|---|---|
| Selection of information source | Data ingestion | Pre-processing | Information extraction | Database | Presentation area |
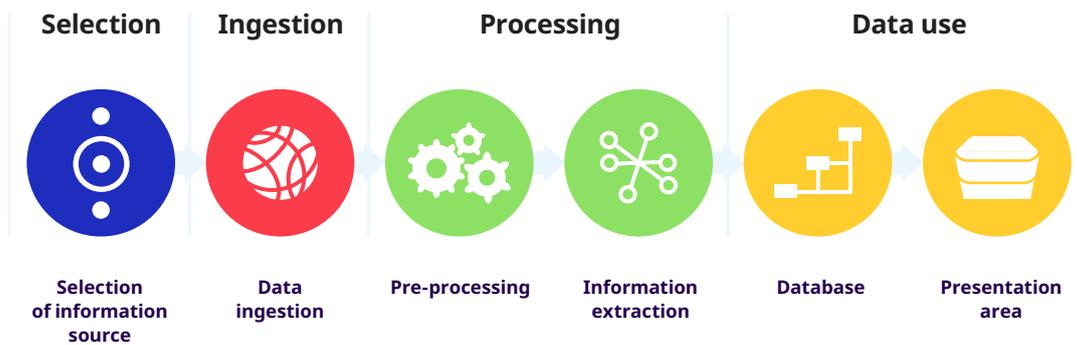
**Source:** based on Cedefop, 2019b.

The Skills OVATE uses three methods of collecting OJVs:

▶ The first and most significant method is direct access via **application programming interface (API)**, which allows downloading of vacancy content directly from OJV portal databases. This direct access requires a formal agreement with the website operator and incurs maintenance and agreement costs. Data collected in this way is of higher quality and can be downloaded much faster than data collected by other methods.

▶ **Scraping** is also used to extract structured data from websites. Web scraping relies on data already being structured on the web page so that it can be accurately extracted by knowing the exact position of each field on the web page. As specific web scrapers must be programmed for each website, this is ideal for sites which contain many vacancies.

▶ **Crawling** is used only as a last resort. This method, which uses a programmed robot to browse web portals systematically and download their pages, is much more general than scraping and is easier to develop. However, crawlers collect significantly more website noise (irrelevant content), so that more effort is needed to clean the data before further processing.

As part of the methodology employed by Cedefop, OJV portal owners were well informed about the intended collection of data and, where possible, written agreements were made to formalize cooperation between Cedefop and the portal owners. In many cases owners granted direct access to data via API. Where this was not possible, scraping and crawling were used. Rejection of access to data sources was negligible (around 0.1 per cent of indexed vacancies) and came mainly from small websites with low market coverage.

Vacancy sources vary in quality and content. To develop a database suitable for subsequent analysis, pre-processing involves the following actions:

▶ **Cleaning:** OJVs are designed to attract the most suitable candidate, not to provide clean data for labour market analysis. Alongside analytically useful information, OJV analysis can also download "noise" (such as advertisements, unticked options in drop-down menus, company profile presentations, etc.). Cleaning involves a sequence of activities to remove "noise" from the data to improve its quality and prepare it for the following phases.

▶ **Merging:** Employers often post a vacancy on more than one portal. Aggregators – webpages collecting vacancies from a number of job portals – increase the chance of finding duplicates. Duplications of vacancies are not desirable for the final analysis, but in the initial part of the pre-processing phase

they can enrich the data by enabling different information on the same vacancy posted in different places to be collected and combined.

▶ **De-duplicating:** Once data for the same vacancy has been merged, it is necessary to remove duplicate vacancies from the analysis. An OJV is considered a duplicate if the description and job location are the same as those in another job advertisement in the database. Vacancy metadata (such as reference ID, page URL) can also be used to identify and remove job vacancy duplicates on aggregator websites.

Information from vacancies is extracted using ontology-based and machine-learning algorithms. Ontologies rely on a pre-existing framework for processing and analysis of online vacancies. Cedefop's system uses both standard ontologies (ESCO for occupations and skills, NACE for industry, NUTS for places of work, ISCED for education level) and custom-based ontologies (by type of contract, experience, salary, working hours) related to skills and the job market. Using the power and flexibility of machine-learning algorithms, the content of job advertisements is matched to available ontology terms, following initial system training. Each vacancy is processed in its original language, which has its own unique logical framework. In the first instance, information in vacancies is classified using text matching and/or similarity. Ontologies are continuously updated and enriched, on the basis of both automatic processes and expert validation. In cases where no results are obtained, information is extracted using a machine-learning algorithm, trained on the basis of large data sets to best fit variables in a given language. In Cedefop's system, machine learning is currently used to classify only occupations; application of machine learning to other variables is under development.

## 1.1.5.  Online dissemination and future work

Skills OVATE offers detailed information on the jobs and skills employers are asking for through OJVs. At present, the tool presents data collected from July 2018 up till March 2020 and cover the EU Member States. Cedefop is also working on expanding the functionalities of the system.

Although the information has been generated using a robust methodology and underlying online job vacancy data are already of high quality, more work is needed to clean and classify the data. More analysis and testing are also necessary to enable Skills OVATE to provide sound evidence to inform policy developments. Nevertheless, Cedefop has taken the decision to publish early information on skills already identified in OJVs in order to show the data's potential, stimulate debate and provide information to help contextualize labour market developments.

While continuing to develop the underlying infrastructure that will permit a fully fledged analysis of the microdata in the future, Cedefop has already made a series of dashboards available for exploration (see figure 1.3). These offer a preliminary glimpse into employers' demand for workers in a range of occupations (up to ISCO four-digit level) and for skills in EU countries/regions and sectors, broken down by occupation, and enable analysts to identify those skills and skill sets most in demand within occupations.

Work in the near future will include continuous efforts to further increase data quality. A key aim is to improve the capacity of the system to extract skills from OJVs in countries where the average number of skills extracted per vacancy (the skills "yield") is rather low. Further work will focus on cooperating with Eurostat to pave the way for using OJVs to deliver reliable statistics at EU level. Cedefop will also engage in more fundamental research to develop new insights in a number of areas, including:

▶ skills, occupations and emerging trends within them that may be used to inform (new and existing) hierarchies and taxonomies;

▶ transferability of skills demanded across occupations and industries;

▶ employer demand for new digital technologies (e.g. artificial intelligence, AI) and skills;

▶ indicators of skills, and occupational imbalances and mismatches;

▶ employer demand for vocational education and training (VET) qualifications and skills;.

Over time, such research will facilitate the development of new types of skills intelligence that can strengthen the evidence base for policy-making in the EU and beyond.

▶ **Figure 1.3.  Disseminating information on OJVs: A sample dashboard from Cedefop's Skills OVATE**
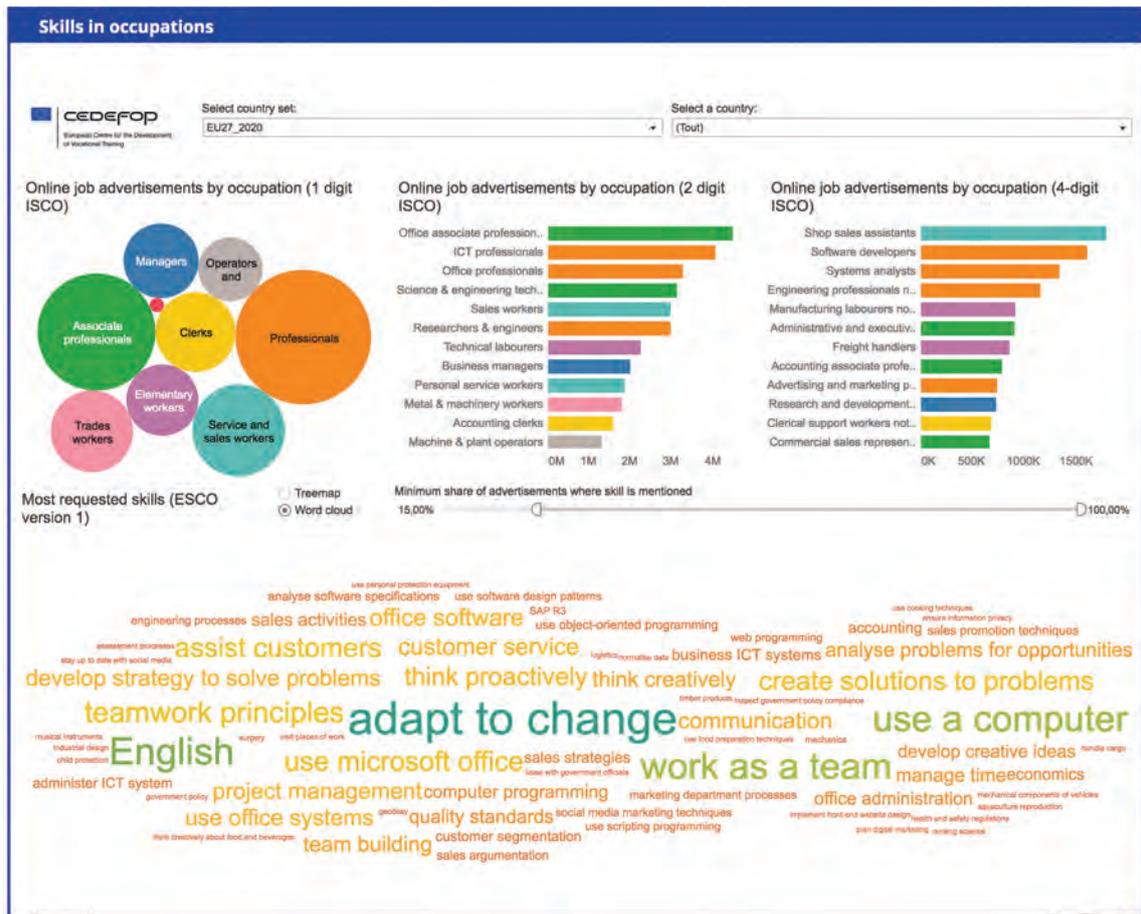


**Source:** https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies. (24 Aug. 2020).

## ▶ 1.2. The European Training Foundation and big data for labour market intelligence: Shaping, applying and sustaining knowledge

**Eduarda Castel Branco**

### 1.2.1. Introduction

Governments and social partners in most partner countries of the European Training Foundation (ETF) are unanimously agreed on the need to develop and make better use of information on labour market and skills dynamics in order to improve the performance of education and training, the availability of qualifications and skills for employment, and people's lifelong societal and personal development.

In this context, most partner countries have been reinforcing their systems, capacities and methods to identify, analyse and anticipate the demand and need for skills in a context of changing economic structures, new types of work, and rapid digitalization of occupations and tasks.

While conventional LMI, based on regularly collected statistics, surveys for specific purposes and qualitative methods, has gained ground in ETF partner countries, there is much room for further innovation in data sources, improvement in analytical capacities, and modernizing of formats and instruments for visualizing and disseminating insights for users, including policy-makers, social partners, and education and training providers.

Big data analytics offers new possibilities for improving LMI and delivering real-time and fine-grained skills analysis and insights for users. Big data is all around us, characterized by "four Vs": volume, variety, velocity and – eventually – value. Machine learning and AI algorithms, combined with immense computing power, allow data science to exploit certain big data sources that have great potential to supplement and enrich conventional LMI. Among these sources are the OJVs managed by a large variety of online job portals and boards.

Creating knowledge out of large volumes of data that are available with high velocity and variety is the major goal of big data analysis. The key concept is that of value. Analysis of thousands and millions of job vacancies can tell us a lot about the skills employers want, in almost real time and in fine-grained detail. Screening and ranking of OJV portals – the first step in the methodology – can inform us about the overall panorama of online LMI in countries and regions, the features of the individual job portals, the volume of posted OJVs, and the sectoral and occupational coverage of those OJVs. Most importantly, analysis of OJVs reveals the specifics of how employers describe jobs and tasks, the mix of skills they seek, the importance they attribute to credentials and qualifications, and the conditions they offer.

### 1.2.2. Big data for LMI: The ETF project

ETF partner countries have seen a growing use of digital tools and online portals – both public and private – to post and manage job vacancies.

In this context, in 2018 the ETF launched an initiative within its digital innovation project aimed at exploring the potential for application of big data analytics for LMI in ETF partner countries. This initiative has brought the ETF together with data scientists and researchers in the Centro di ricerca interuniversitario per i servizi di pubblica utilità (Interuniversity Research Centre on Public Services, CRISP)[4] at the University of Milan-Bicocca (see Mezzanzanica and Mercorio, 2019a, b).

The project approach entails the combination of four main elements, set out schematically in figure 1.4.

---

4   http://www.crisp-org.it/ [13 Feb. 2020].

▶ **Figure 1.4. The ETF initiative on big data for LMI: The four main elements**



**1.**
**Methodological**
**guidance**

**2.**
**Application**
**begins**
**in partner**
**countries**

**3.**
**Expert**
**network**

**4.**
**Knowledge-**
**sharing**
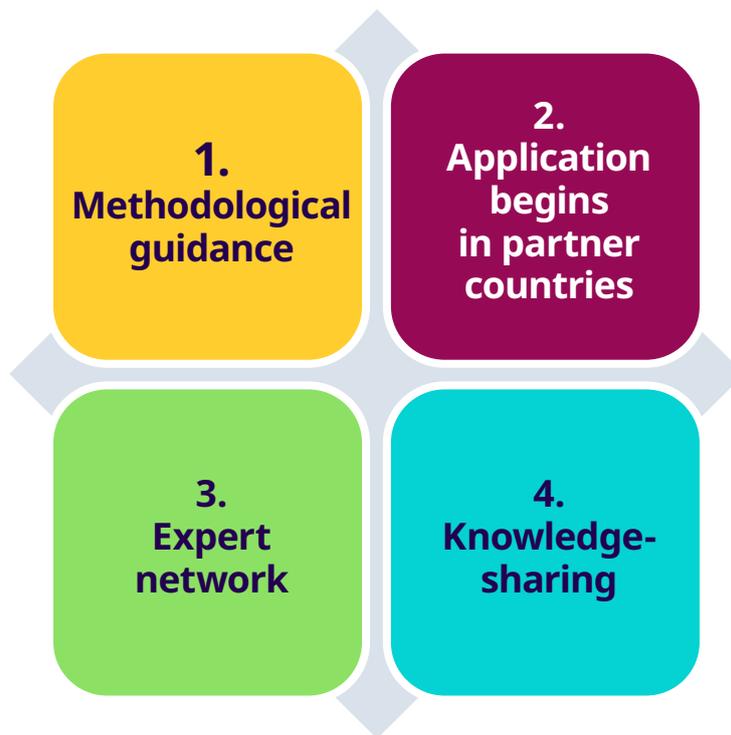
**Source:** Author.

## 1. Methodological guidance (2019)

This first stage entailed creating a knowledge base on the application of big data for LMI (demand for jobs, qualifications and skills), with the aim of providing guidance to researchers, statisticians and those involved in making decisions on the establishment of big data models for LMI. This resulted in the publication in 2019 of a short handbook entitled Big data for labour market intelligence: An introductory guide (Mezzanzanica and Mercorio, 2019a).

Aimed at statisticians, researchers, policy analysts and decision-makers in the ETF's partner countries who face the challenges of anticipating and disseminating  insights into the dynamics of demand for jobs, skills and qualifications, this handbook addresses key conceptual, methodological and organizational aspects of using big data for LMI. It clarifies the ways in which big data can be used to transcend the limits of conventional approaches to LMI and add value to established statistics.

## 2. Prototype OJV analysis system with data and visualization tool (2019–21)

In this second stage, the methodology outlined in the handbook is being applied in a selected number of ETF partner countries, structured according to the five steps of the "knowledge discovery in databases" (KDD) process (see Fayyad, Piatetsky-Shapiro and Smyth, 1996; also section 1.2.4 below), and taking place in three phases (see figure 1.5):

▶ phase 1, 2019: application of methodology begins in Tunisia and Morocco with a landscaping and feasibility study report on the web labour market;

▶ phase 2, 2020: launch and full application of methodology in two pilot countries, Ukraine and Tunisia, and expansion to new countries to be selected;

▶ phase 3, 2021: follow-up actions for sustainability in the countries involved.

▶ **Figure 1.5:  Data flow, from data collection to results presentation**

## Overall data flow



| Ingestion | Processing | | | Front end |
|---|---|---|---|---|
| Data collection | Pre-processing | Classification | Post-processing | Presentation area |
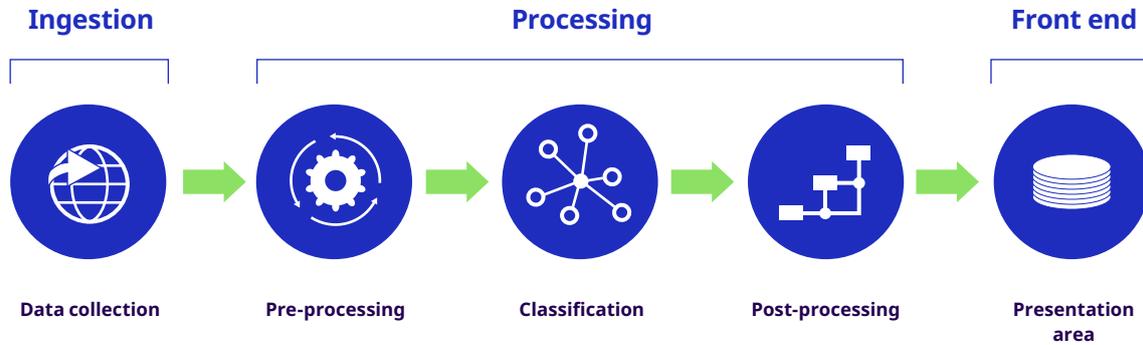
**Source:** CRISP, technical meeting, ETF, 20 May 2019.

## 3. Network of experts (from 2020 as part of the ETF Skills Lab)[5]

Turning big data into LMI requires multidisciplinary expertise, notably in:

▶ statistics;

▶ technical areas such as data science, data analysis, software development;

▶ labour markets (demand and supply).

A network of experts dedicated to sharing knowledge and developing the necessary skills for the application of big data analytics in LMI can be developed along the lines summarized in figure 1.6.

▶ **Figure 1.6.  Main features of a network of experts to develop big data for LMI**



| Aims | Learning, sharing, managing knowledge | Peer support in planning and implementing big data LMI projects | Sustainability and convergence with statistical offices |
|---|---|---|---|
| Work plan and steering | Specific network objectives: 1-3 years | Mapping existing practice big data, leading institutions, use of results (2020) | Steering committee, chair, secretariat |
| Work modalities | Structured work-shops: peer learning, peer review (incl. online); conferences | Publications; repository | Part of Skills Lab: web presence |
| Composition | Multi-disciplinary expertise mix | Wide coverage of ETF partner countries, gradual growth | Interested institutions, other regions |

**Source:** Author.

5   See https://www.etf.europa.eu/en/projects/skills-lab [27 July 2020].

The establishment of a balanced mix of disciplines and areas of expertise will build on:

▶ existing expert networks involved in ETF activities, such as "Make it Match" (Eastern Partnership);

▶ contacts of ETF country coordinators and ETF thematic experts;

▶ research and literature;

▶ results of good practices in use of big data LMI.

The network can begin with an initial core membership and then gradually expand to complete its programme of activities.

## 4. Knowledge-sharing and skills development (2019–21)

The multi-disciplinary network of experts will initiate and pursue knowledge-sharing and skills development activities with the following objectives:

▶ developing new knowledge on changing skills demand;

▶ sharing a common set of concepts, principles, knowledge and good practice;

▶ updating and further developing the methodological guidance, building on Mezzanzanica and Mercorio, 2019a;

▶ stimulating the multi-disciplinary approach within country and cross-country teams;

▶ identifying areas and themes requiring more and deeper skills development within the expert network, and identifying ways forward for training;

▶ conducting a targeted training programme(s) on the application of big data analytics for LMI, on the visualization and interpretation of results, and on state-of-the-art architectures, technologies and tools, to be accompanied by assessment of learning outcomes, and the award of a validation document signed by the trainers and the ETF;

▶ supporting the sharing of thematic information and questions and answers via an online platform (possibly, but not necessarily, the ETF Skills Lab or another).[6]

Alongside the regular members of the network, other interested participants can attend its activities, notably the training elements.

## 1.2.3. Questions and reflections

The practical application of the ETF project began in 2019 (phase 1 above) with the screening and ranking of 16 online job portals in Tunisia and 15 in Morocco; follow-up steps are under way. This experience can shed light on the specifics of using big data analytics for LMI in countries of the European neighbourhood. Key themes and questions for consideration include the following:

▶ Cooperation with statistical offices: In what conditions can real-time LMI be used by statistical offices? Can real-time LMI supplement and enrich labour market statistics?

▶ Cooperation with national/regional research institutes and agencies: Can multi-disciplinary teams be established combining data science with labour market expertise? How well developed are multi-disciplinary approaches?  Are there data scientists and data engineers with the required skills available in partner countries?

▶ Creating trust and cooperation with OJV portals for data sharing and use: Which approaches are most suitable?

▶ Are the volumes of OJVs sufficient to create real-time LMI in most countries?

---

6   https://www.etf.europa.eu/en/projects/skills-lab [14 Aug. 2020].

▶ Is cross-country comparability of results on skills demand useful? If so, under what conditions, and which comparable classifications can be used?

▶ Are there potential benefits to be reaped in cooperating across countries and regions in developing real-time LMI (e.g. establishing some common principles and methodological instruments and common visualization platforms, and sharing knowledge) and in creating a sound basis for collaboration with statistical offices?

▶ Removing the "mystery" and "information barriers" surrounding big data analytics and the role of machine learning: this may involve, for example, intellectual scrutiny, new angles of interpretation, and the provision of information for specialists and managers in relevant institutions.

▶ Systematic cooperation and exchanges, notably with groups and organizations at the leading edge of big data analytics, such as Burning Glass Technologies and Cedefop.

## 1.2.4.  A methodology for turning big data into LMI[7]

Figure 1.7 presents in schematic form the five-step KDD approach to turning big data into LMI.

▶ **Figure 1.7.  The KDD process, showing the "big data Vs" involved in each step**
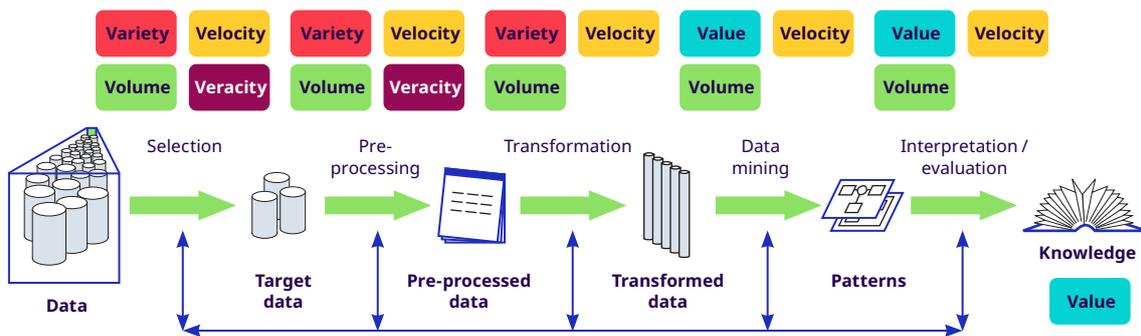


**Source:** Mezzanzanica and Mercorio, 2019a.

## Step 1: Selection

The first step is the selection of data sources. Each web source needs to be evaluated and ranked in terms of the reliability of the information it offers. This process should take into account, for example, the vacancy publication date, the frequency with which the website is updated, the presence of structured data and any downloading restrictions. At the end of this phase, a *ranking* of reliable web sources is produced. This step involves all of the big data "four Vs" mentioned in section 1.2.1 above, along with a fifth, *veracity*, that is, the degree to which the data is free from biases, noise and abnormalities.

Key questions posed by the selection phase for the various members of the expert network include the following:

1. For the statistical experts: How do we identify the criteria to be included in the source model, and how do we extract these criteria (variables) from the sources? How do we rank the sources?

2. For the technical experts: How do we identify an appropriate data model paradigm (e.g. relational, document, key-value, graph, etc.) to store huge volumes of data? How do we collect data automatically? Do we need API access, or do we need to develop a web scraper/crawler? How do we schedule automatic data collection processes?

3. For the labour market domain experts: How do we select the right sources? Have we selected the right sources?

---

7   This section is an edited extract from Mezzanzanica and Mercorio, 2019a.

## Step 2: Pre-processing

This step includes data cleaning to remove noise or inappropriate outliers (if any), deciding how to handle missing data, and identifying a way to detect and remove deficient entries (e.g. duplicated vacancies and vacancies with missing values). Data quality and cleaning are essential tasks in any data-driven approach to decision-making, in order to guarantee the soundness of the overall process, i.e. the extent to which data is accepted or regarded as true, real and credible (see e.g. Boselli et al. 2015; Redman, 1998; Wang and Strong, 1996).

Identification of duplicated job vacancies is far from straightforward. The posting of the same job vacancy on multiple websites (which is usual practice) is duplication; re-use of the *same* text to advertise a *similar* position, however, is not. Identification of appropriate features to enable correct recognition of duplicates is crucial in the web labour market domain. The pre-processing step reduces the complexity of the big data scenario, improving the *veracity* dimension through data quality and cleaning.

Key questions raised by this second phase for the network of experts include the following:

1. For the statistical experts: How do we evaluate data consistency? How do we measure data accuracy? How do we estimate data significance?

2. For the technical experts: How do we identify duplicate data records? How do we identify missing values?

3. For the labour market domain experts: How do we identify synonyms used in the labour market domain and thereby contribute to improving data accuracy? How do we identify criteria that characterize missing values and duplicates?

## Step 3: Transformation

This step includes data reduction and projection, with the aim of identifying a unified model to represent the data, depending on the purpose of the exercise. It may also include a reduction in the number of dimensions or transformative methods to reduce the effective number of variables or to find invariant representations for the data. Like step 2, the transformation step reduces the complexity of the dataset, in this case by addressing the *variety* dimension. It is usually performed by means of "extract, transform, load" (ETL) techniques, which support steps 2 and 3 in the KDD process. Roughly speaking, through ETL, the data extracted from a source system undergoes a series of transformation routines that analyse, manipulate and then clean the data before loading it into a knowledge base. By the end of this step, whose outcome is a clean, well-defined data model, the big data *variety* issue should be resolved.

Key questions raised by the transformation phase for the various members of the expert network include the following:

1. For the statistical experts: How do we measure the completeness of the target model identified? Does the target model maintain data significance at the end of the ETL process?

2. For the technical experts: How do we develop big data procedures to transform raw data into a target model in a scalable manner?

3. For the labour market domain experts: How do we identify the destination data format and taxonomy?[8]

## Step 4: Data mining and machine learning

The aim of this step is to identify appropriate types of AI algorithms (e.g. classification, prediction, regression, clustering, information filtering) to serve the purpose of the analysis, by searching for patterns of interest in a particular representational form. In the specific context of LMI, text classification algorithms (based on ontology, machine learning, etc.) will usually be required, in order to build a classification function for

---

8 The labour market is characterized by several standard taxonomies, such as ISCO/O*NET/SOC for occupations, ESCO for skills, NACE for classification of economic activities, etc.

mapping data items into several pre-defined classes. This step is a particularly crucial one, as it is mainly devoted to the extraction of knowledge from the data.

Key questions raised by this phase for the network of experts include the following:

1. For the statistical and technical experts:  How do we select the best algorithms? How do we tune their parameters? How do we evaluate algorithm effectiveness? How do we implement it at scale?
2. For the labour market domain experts: Which knowledge should be selected and which should be discarded? What is the labour market significance of the knowledge obtained? What novel insights have been discovered in this way? How do we explain the results of the mining process from a labour market perspective?

## Step 5: Interpretation/evaluation

This final step employs visual paradigms to portray the knowledge obtained in ways that serve the user's objectives. In the LMI context, this means taking into account the user's ability to understand the data and their main goal in the LMI field. For example, government agencies might be interested in identifying the most sought-after occupations in particular local areas; companies might focus on monitoring skills trends and identifying new skills in demand for certain occupations, so that they can design training paths for their employees. In the past few years, much work has focused on producing off-the-shelf visual libraries, based on a variety of narrative and visual paradigms.

Key questions raised by this final phase for the network of experts include:

1. For the statistical and technical experts: How do we select the visualization paradigm? How do we select an appropriate visualization model for the knowledge we want to portray?
2. For the labour market domain experts: How do we deliver appropriate knowledge according to stakeholder needs? How do we identify visual navigation paths for each stakeholder? How do we retrieve feedback (if any) from labour market users? How do we insert labour market knowledge into the business environment?

As may be observed from this summary of the five-step KDD process, the number of technical and statistical issues decreases as the process advances, while the number of issues and challenges facing the labour market experts increases.

# ▶ 1.3. Can we use big data for skills anticipation and matching? The case of online job vacancies

**Fabio Mercorio and Mario Mezzanzanica**

## 1.3.1. Quo vadis labour market?

Over recent decades, significant forces and factors have dramatically changed the nature and characteristics of the labour market in both advanced and developing countries. Technical progress, globalization and the reorganization of the production process – with outsourcing and offshoring – have radically altered the demand for a range of skills. At the same time, population ageing in advanced economies has intensified the need for continuous training and is also likely to increase the structural demand for certain competencies, such as those related to the health and care of the elderly. The overall impact of these factors on the labour market is multi-faceted. On the one hand, while some jobs are disappearing, others are emerging; of the latter, some are simply variants of existing jobs, while others are genuinely new jobs that did not exist until a few years ago. On the other hand, the skills and qualifications associated with the new labour market have changed substantially, in both quantity and quality. It is not simply that new skills are needed to perform new jobs; the skill requirements of existing jobs have also changed considerably.

This evolution in the labour market raises many questions, for example:

▶ How are we to capture and understand the changes that are taking place?

▶ Which occupations will grow in the future and where?

▶ What skills will be most in demand in the years ahead?

▶ Which skills should workers and jobseekers learn to fit into the evolving labour market better?

▶ Which jobs are more likely to be affected by digitalization and robotization, and how are we to estimate this impact?

▶ How are employers to be sure that the skills held by people working in their organizations are adequate to meet the demands of today's labour market?

To address these questions, the application of AI to OJVs has recently been proposed as a way to process and disclose the information embedded in the online labour market to support the decision-making activities of labour market specialists through LMI (see e.g. UK Commission for Employment and Skills, 2015; Mezzanzanica and Mercorio, 2019c). Though there is no unified definition of what LMI is, in the present context we may take it to refer to the design and use of AI algorithms and frameworks to analyse labour market data to support decision-making. In such a context, LMI represents added value for both public and private operators seeking a deep understanding of labour market dynamics, occupations, skills and trends. The advantage of this approach includes its capacity to

▶ monitor labour market dynamics and trends in near real time, then generate practically useful information much more rapidly than classical survey-based analyses, enabling an immediate response to labour market expectations in terms of both skills and occupations;

▶ observe the labour market over several dimensions (e.g. territory, sectors, contracts, salary, digital/ soft skills, etc.) at a very fine-grained level;

▶ evaluate and compare local and international labour markets to support fact-based decision-making for skills anticipation and matching, organizing of training, and upskilling/reskilling activities based on labour market demands.

## 1.3.2.  LMI and big data: Current work and future potential

The challenge of monitoring, analysing and understanding these labour market changes both promptly and  in fine-grained local detail is now one of great practical significance, and various responses are already under way. Machine learning has been applied to compute the effect of robotization within occupations in the US labour market (Frey and Osborne, 2017), as well as to analyse skill relevance in the US standard taxonomy O*NET (Alabdulkareem et al., 2018), to cite just two recent initiatives. In 2016, the EU and Eurostat launched the ESSnet big data project,[9] involving 22 EU Member States, with the aim of integrating big data into the regular production of official statistics. In 2014, Cedefop invited tenders for realizing a system able to collect and classify OJVs from five EU countries (Cedefop, 2014). Since then, a further project has been launched to scale up the machine-learning-based system and extend it to the whole EU, including all 27 Member States and all 32 languages of the Union (Cedefop, 2016). The rationale behind all these projects is to turn data extracted from OJVs into knowledge (and thus value) for policy design and evaluation through fact-based decision-making.[10]

What big data actually is, what makes data "big" and what does not, as well as the challenges and opportunities involved in dealing with big data, are all questions that are still open to debate.[11] Nevertheless, there is undeniably a growing interest in the manipulation and use of big data, driven by the potential it offers managers to monitor their businesses more effectively, thereby acquiring knowledge that can improve their decision-making and performance.

Moreover, the knowledge resulting from the more detailed classification obtainable from analysis of OJVs structured by standard taxonomies such as ESCO, NACE etc., and from the ability to extract information about skills, provides the basis for a wide range of analyses that can be used to explain phenomena as well as to enrich understanding of labour market dynamics. In the next two sections we present some examples.

## 1.3.3.  Identifying new (potential) emerging occupations

The labour market knowledge generated by analysis of OJVs can be used to identify new (potential) emerging occupations, by exploiting unsupervised learning techniques. The term "new (potential) emerging occupations" refers to occupations that have not yet been included in the system of occupational classification. Clearly, the use of a new term in advertising a job does not itself identify a new occupation, as this new emerging term has to be validated by a growing trend over time that confirms the establishment of a new (emerging) occupation in the online labour market.

The main idea behind the use of AI algorithms in this context is to compute the language that characterizes each occupation code through the lexicon used within the vacancies by means of word-embeddings. This is done by vector representation of words – a process that belongs to the family of neural language models (Bengio et al., 2003) – where every word of the lexicon is mapped to a unique vector in the corresponding $n$-dimensional space. After Word2vec training on the lexicon, words with similar meaning are mapped to similar positions in the vector space. For example, "powerful" and "strong" are close to each other, whereas "powerful" and "Paris" are further apart. The word vector differences also carry meaning (Mikolov, Yih and Zweig, 2013). This approach allows specific words to be positioned in the n-dimensional space, enabling us to compute the similarity between job vacancies in terms of both titles and skills content.

The idea of the approach is depicted schematically in figure 1.8. We begin with classified OJVs (for a description of the selection and classification process, see section 1.2.4 above). Then we build up several vector-space representations of words (describing occupations and skills) to catch lexical similarities between OJVs. The system then computes similarities between known terms (occupations and skills) and new ones; and finally, the system suggests new potential occupations for human or AI validation.

---

9   See https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data [27 July 2020].

10   Some results have recently been published: see Boselli et al., 2017, 2018.

11   For a detailed discussion on the use of big data for LMI, see Mezzanzanica and Mercorio, 2019a.

▶ **Figure 1.8. Detecting new (potential) emerging occupations through AI**
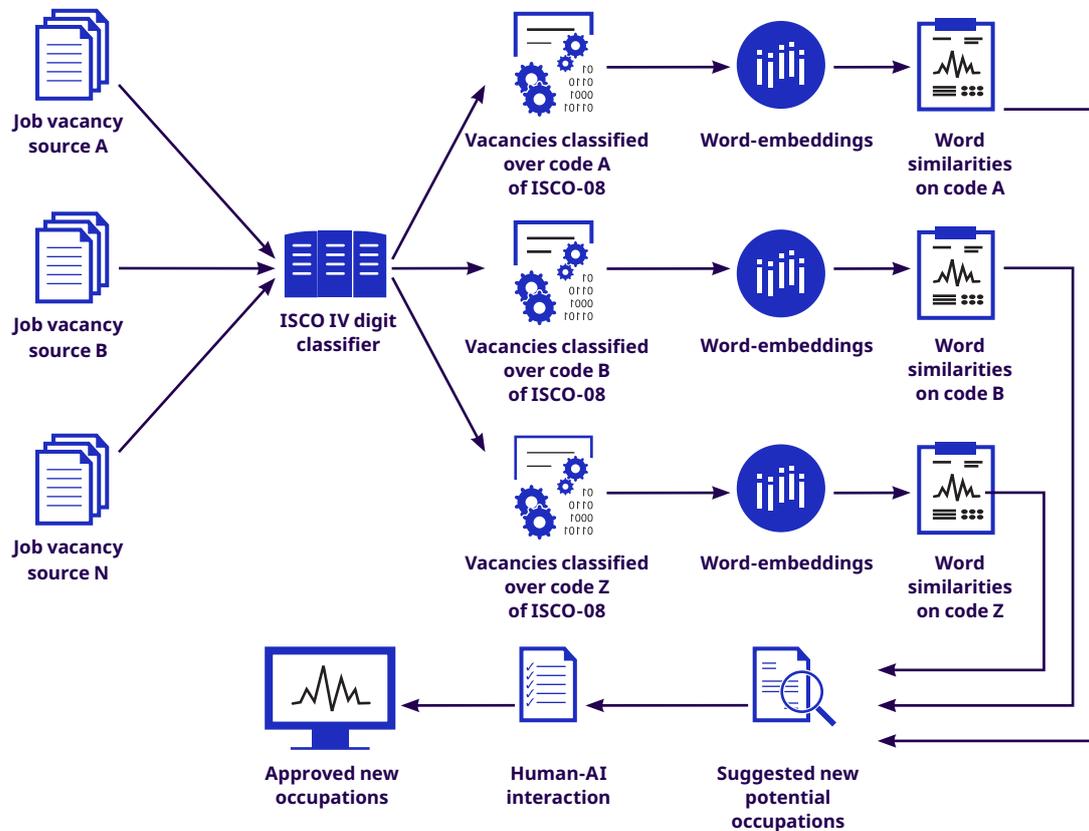


**Source:** Mezzanzanica and Mercorio, 2019a.

The classification of OJVs and the identification of new (potential) emerging occupations has been applied to the Italian labour market to build up an online digital skill observatory[12] that focuses on ICT-related jobs, as shown in figure 1.9. Note that:

1. OJVs and skills are classified according to the European e-Competence Framework (e-CF),[13] structured around four dimensions. These dimensions reflect different levels of business and HR planning requirements in addition to job/work proficiency guidelines. Figure 1.9 shows the first dimension, derived from the ICT business processes PLAN – BUILD – ENABLE – MANAGE – RUN. Specifically, PLAN, BUILD and RUN are core areas, while ENABLE and MANAGE are cross-cutting issues related to the core areas.

2. An asterisk against an occupational title in the figure indicates a new (potential) emerging occupation, not included yet in the 30 European ICT professional role profiles of the e-CF framework, and identified as discussed above. This allows analysts and labour market specialists to analyse and compare both (i) traditional and consolidated occupations, and (ii) potential novel occupations as a whole.

3. A fine-grained analysis of each occupation can be done to arrive at a more detailed comparison of the skills that are most significant in each profile, as well as comparison across sectors and territories.[14]

---

12 The project has been funded by the Italian ICT unions, and realized by Tabulaex and CRISP-Unimib Research centre. See https://competenzedigitali.org/ (Italian only) [13 Feb. 2020].

13 Specifically, OJVs are classified using the 30 European ICT professional role profiles built on the e-CF, while skills are classified using the e-CF competence framework. For details, see https://www.ecompetences.eu/it/ict-professional-profiles/ [13 Feb. 2020].

14 See http://competenzedigitali.org [13 Feb. 2020].

▶ **Figure 1.9.  Distribution of ICT-related OJVs in Italy in 2018, classified using the e-CF standard**

| Profession | Plan | Build | Enable | Manage | Run |
|---|---|---|---|---|---|
| **Professions (CENª)** | 19% | 30% | 16% | 26% | 8% |
| Account manager | 6% | 14% | 38% | 33% | 9% |
| Artificial intelligence specialist* | 14% | 23% | 20% | 36% | 7% |
| Big data specialist* | 17% | 37% | 18% | 23% | 5% |
| Blockchain specialist* | 17% | 19% | 19% | 37% | 8% |
| Business analyst | 13% | 21% | 28% | 33% | 5% |
| Business information manager | 13% | 23% | 27% | 31% | 6% |
| Chief Information Officer | 11% | 26% | 19% | 34% | 10% |
| Cloud computing specialist* | 14% | 25% | 20% | 32% | 8% |
| Data scientist | 14% | 32% | 23% | 26% | 4% |
| Data specialist | 13% | 26% | 21% | 33% | 6% |
| Database administrator | 23% | 33% | 13% | 22% | 9% |
| Developer | 24% | 35% | 12% | 22% | 7% |
| Development Operations Expert | 16% | 32% | 15% | 28% | 7% |
| Digital consultant | 14% | 18% | 22% | 38% | 8% |
| Digital educator | 11% | 14% | 21% | 44% | 10% |
| Digital media specialist | 13% | 18% | 34% | 29% | 5% |
| Digital transformation manager | 8% | 13% | 38% | 36% | 5% |
| Enterprise architect | 16% | 31% | 16% | 29% | 8% |
| ICT operations manager | 8% | 17% | 26% | 40% | 10% |
| Information security manager | 10% | 22% | 20% | 41% | 8% |
| Information security specialist | 12% | 21% | 18% | 43% | 7% |
| Internet of Things specialist* | 21% | 27% | 14% | 28% | 9% |
| Mobile specialist* | 24% | 35% | 11% | 24% | 5% |
| Network specialist | 19% | 30% | 15% | 24% | 12% |
| Product owner | 16% | 18% | 20% | 42% | 5% |
| Project manager | 8% | 20% | 20% | 46% | 6% |
| Quality assurance manager | 14% | 21% | 15% | 41% | 9% |
| Robotics specialist* | 26% | 21% | 13% | 33% | 7% |
| Scrum master | 11% | 20% | 20% | 43% | 6% |
| Service manager | 9% | 17% | 24% | 39% | 11% |
| Service support | 9% | 24% | 17% | 27% | 23% |
| Solution designer | 17% | 25% | 18% | 32% | 8% |
| Systems administrator | 16% | 37% | 10% | 23% | 14% |
| Systems analyst | 16% | 33% | 13% | 26% | 12% |
| Systems architect | 16% | 28% | 17% | 30% | 8% |
| Technical specialist | 12% | 29% | 16% | 25% | 18% |
| Test specialist | 17% | 23% | 15% | 35% | 9% |

▶ Plan    ▶ Build    ▶ Enable    ▶ Manage    ▶ Run

**Note:** ª CEN = European Committee for Standardization    * indicates a new (potential) emerging occupation.
**Source:** Italian Observatory of Digital Competences: see http://competenzedigitali.org.

## 1.3.4. Hard/soft/digital skills rates

The prominence of ICT in everyday life, as well as the pervasiveness of ICT solutions and tools in the modern labour market, suggest the value of a deeper investigation of the skills related to ICT, aimed at establishing the relevance of these skills to a wide range of occupations. A top-down approach to this question, such as those linked to surveys and questionnaires, would not allow us to obtain precise information. OJVs, by contrast, offer the possibility of applying a bottom-up approach and calculating a digital skills rate (DSR) to measure the incidence of digital skills within occupations, whether these are directly related to the ICT world or not. Skill rates would indicate the prominence of soft and hard, digital and non-digital skills within each occupation.

Using this approach, we can compute the DSR for each ESCO occupation (four-digit level) in each EU country, and at a territorial level using the NUTS taxonomy (e.g. NUT1 for macro-regions).

Furthermore, given the relevance of digital skills in today's labour market, we can augment the ESCO classification by further constructing the following sub-groups:

▶ **information brokerage skills**, i.e. the ability to use ICT tools and platforms for data exchange and communication (e.g. on social media);

▶ **basic ICT skills**, i.e. the ability to use some standard ICT applications in support of individual professional activities (e.g. use of spreadsheet or word-processing software);

▶ **applied/management ICT skills**, i.e. the ability to use tools and software to support management, operational and decision-making processes within the organization (e.g. administrative software);

▶ **ICT technical skills**, i.e. the ability to use solutions, platforms and programming languages that are strongly related to ICT-specific professions (e.g. programming languages, advanced ICT software).

Clearly, this approach can be easily applied to compute both soft skills rates and hard/non-digital skills rates. As a result, the analyst will obtain a map that portrays the skills composition of each occupation at a very fine-grained level.

An example is provided in figures 1.10–1.12, focusing on non-ICT related jobs. First, occupations can be analysed focusing on the sectoral dimension (figure 1.10), comparing how the DSR has evolved between 2014 and 2017, and monitoring both hard non-digital and soft skills rates. Then, once a particular sector has been selected (e.g. services), the fine-grained analysis allows a breakdown of digital skills within the DSR according to the four categories described in figure 1.11. Finally, for each occupation, we can observe the level of each elementary ESCO skill along with the relevance of that skill that emerges from the online labour market (figure 1.12).

▶ **Figure 1.10.  Analysis of DSR by sector (level 1)**



**Source:** Italian Observatory of Digital Competences: http://competenzedigitali.org.
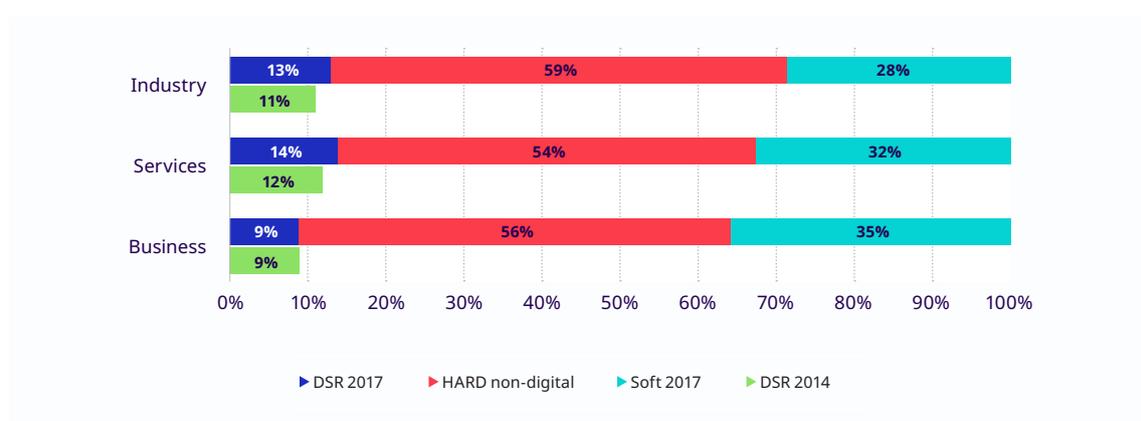
▶ **Figure 1.11.  Analysis of DSR by digital skills (level 2)**



**Source:** Italian Observatory of Digital Competences: http://competenzedigitali.org.
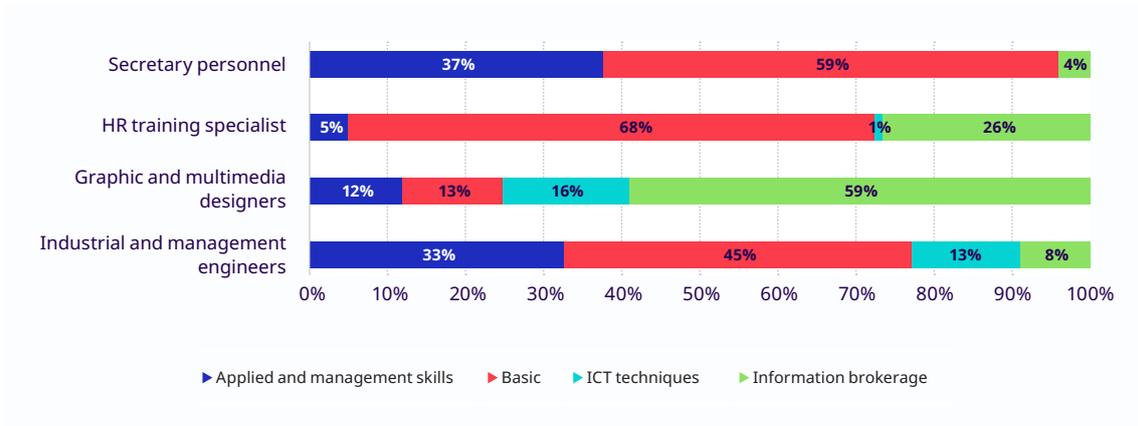
▶ **Figure 1.12.  Analysis of DSR by occupation and elementary ESCO skills (level 3)**

| Occupation | Applied and management skills | | | ICT techniques | | Information brokerage | |
|---|---|---|---|---|---|---|---|
| | Database usage | Programs for draughts man | 3D modelling | Front-end Website implementation | Web programming | Graphic Software Usage | SW markup usage |
| Graphic and multimedia designers | 2.5 | 3.5 | 3 | 4.5 | 2 | 5 | 2 |

| Occupation | Applied and management skills | | Information brokerage | | |
|---|---|---|---|---|---|
| | Database usage | ERP | Digital data management | SEO Search Engine optimization | Social Network Usage |
| HR training specialist | 4.5 | 4 | 4.5 | 4 | 2.5 |

**Source:** Italian Observatory of Digital Competences: see http://competenzedigitali.org.

## 1.3.5. Further research directions

The analysis of OJVs allows us to monitor and understand real-time changes in the online labour market in a timely fashion, at a very fine-grained geographical level, and focusing on skills that are particularly relevant at the time. Actual research and projects undertaken to date have focused mainly on the following areas:

investigating whether a correlation exists between the demand for hard/soft skills and the probability that a job will be replaced by computerization (Colombo, Mercorio and Mezzanzanica, 2019);

enriching a skills taxonomy by using AI techniques to analyse references to occupations and skills in the online labour market;

building up a graph database that encodes all the LMI information for Europe, acting like an extended ESCO taxonomy reflecting the real labour market;

improving the classification process through explainable AI algorithms to make a black-box classification algorithm transparent to end-users.

# ▶ 1.4. From big data to smart data: The misconception that big data yields useful predictions on skills

**Stefan Winzenried**

## 1.4.1. Introduction

In what follows, I will demonstrate the following:

▶ why the collection of skills and other occupational data from big data does not result in creating high-quality smart data;

▶ why implicit information is at least as important as explicit, readily available information;

▶ how skill sets can be accurately modelled for job markets in both emerging and advanced economies;

▶ why it is mainly experts – and not algorithms – that will continue to be responsible for such data modelling in the future.

It has always been a priority of JANZZ.technology to present in-depth and sound views on the topics that are currently attracting most attention. At present, these include big data and skills prediction, reskilling and upskilling, AI and skills matching. Globally, there is extensive discussion about the future of work and the associated challenges, and while much of what is written sounds very inspiring and even reasonably plausible, our research indicates that the reality is much more challenging.

In particular, job and skills matching is a significant topic with far-reaching implications for many new technologies in this field. It seems, however, that to date only JANZZ.technology has been able to prove an ability to provide explicable AI for accurate skill and profile matching. Many applications, including those by major platforms and providers of applicant tracking system and HR software, can rarely meet even modest expectations and appear to have made little progress in recent years. There is much marketing hype in this area, and we need to explore the underlying data set in detail to understand whether performance is actually delivered in real-life applications.

In the course of our research we came across a remarkable article (Cappelli, Tambe and Yakubovich, 2019), from which I would like to quote the following passages:

> "When you talk to HR practitioners who see their colleagues in finance and marketing using these technologies with so much success, part of the question they ask is, why does it seem so hard for us? I think part of the point we wanted to make is that there are systemic and structural differences for HR that do make it harder, when you are building an AI-based system, you need to know what the right answer is, what a good employee looks like, and defining that in HR is already a difficult thing to do."

> "As much as I would like to have people think that we're like wizards with this stuff, that we go into the computer, we take this data and then magic comes out, it's an awful lot of just plugging away."

The two quotes highlight the specific challenges faced by both HR and PES. Big data and AI still have such a hard time today in gaining traction in the field of HR and employment because the quality and explanatory power of methods and results in these fields are limited and lack resilience. We recommend this article to everyone who is interested in AI, big data and HR technology, and even more to everyone who is concerned with predictions about skills, employment markets and the future of work in general.

In general, it should be noted that to date, predictive analysis of all types that is based on big data and determined by a large number of variables is very rarely accurate. The longer the time horizon and the more

variables are included, the less likely it is that such predictions will be completely or even partially borne out in reality. There are some spectacular examples of this, such as the long-term weather forecasts for the summer of 2019 in Switzerland, or the attempt to predict the 2018 football World Cup winner, and many more. This is why such forecasts should carry very big question marks and be used only with the utmost caution.

▶ **Figure 1.13.  Long-term weather forecast for Switzerland, summer 2019, using comprehensive big data models, compared to actual measured temperatures**



▶ Forecast AccuWeather      ▶ Measured values MeteoSwiss
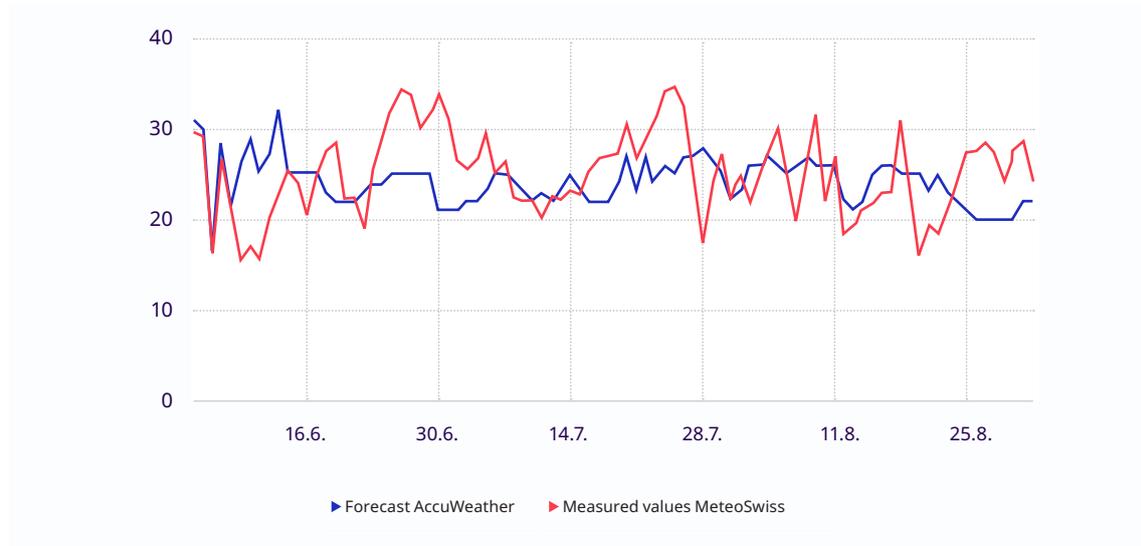
**Sources:** Forecast: AccuWeather; measured values: MeteoSwiss.

The comparison set out in figure 1.13 shows that not even the trend could be predicted with anything approaching accuracy. Long-term weather forecasts are typical examples of large (historical) data models with high variability and limited potential for extrapolation to the future. The forecast accuracy is on average below 10 per cent, a value that would of course not stand up to serious scientific review. Owing to similar very high variability, the same can also be expected to apply for long-term skills forecasts based on big data.

## 1.4.2.  Over ten years of unique experience with occupational big data

At JANZZ.technology we have been dealing exclusively with occupational big data since 2008. For over ten years we have been systematically modelling and curating occupational data worldwide, including skills and competencies, soft skills and qualifications, education and training, certifications and authorizations, in over 40 languages. From the very beginning, the aim was to form big data into smart data using a structured and fully semantic ontological approach. We are convinced, and have repeatedly proven, that this is the only approach that can work reliably in the area of skills matching, as well as in skills anticipation and prediction.

Over the past ten years, we have been largely alone in our approach and our chosen methods. Academia, developers, the ICT community, HR and public employment stakeholders alike have assumed that all their tasks and complex problems could be solved with sufficiently sophisticated algorithms and ever higher computing power, such as cloud computing and big data. It was only around 2015 that competitors such as Google followed our example and began to pursue the ontological approach, i.e. finding solutions by means of knowledge representations.[15]

---

15  For publications and posts that have contributed to this change of opinion in the industry and in academia, see e.g. Winzenried and Zanoli, 2016; Winzenried, 2018; Wissner-Gross, 2016.

Those who have closed their eyes to these developments and have exclusively pursued machine-learning approaches have made little progress over the past five years. At best, their progress has stagnated; some of them have entirely disappeared from the scene.

We are now in a position to say that anyone who works with big data in occupational evaluations, skills matching and skills predictions, and does so without a multilingual and sophisticated knowledge representation with all the necessary semantic dimensions, represents outdated methods that have produced repeatedly disproved results with little to no significance. As is frequently remarked: "A fool with a tool is still a fool."

Typical examples of this type of work are approaches that compile and evaluate all available job advertisements from all available sources in a market over a period of years, and use these aggregated data to make recommendations for market participants, including forecasts of the future employability and required skills of jobseekers.

In the remainder of this section, I will analyse this methodology and point out the shortcomings of the results it produces.

## 1.4.3.  The importance and definition of skills and competencies: A brief examination

What do we mean by the terms "skills", "abilities" and "competencies"? Definitions diverge widely depending on whom you ask. Some people use "skills" to mean what other people understand by "soft skills". Competencies are another topic altogether. For example, at O*NET you can find knowledge and technology skills and tools, which are taken to refer only to directly job-related or transferable knowledge. Why do these terms and expressions overlap significantly? If you want to think about the skills which will be required in the future and make forecasts about how these requirements will change (which skills will gain in importance and which skills will become obsolete), or if you want to perform target-oriented skills matching, you first have to be able to correctly recognize, understand, assign and classify today's skills. Most approaches and big data evaluations fail miserably at this, and hence cannot be used for matching.

Thus, although the same representations and forecasts are repeatedly produced, if they are checked retroactively for accuracy it is apparent that most of them cannot fulfil even very modest expectations.

▶ **Table 1.2.  Typical forecasts of general "top skills", as published regularly by LinkedIn and the World Economic Forum**

| Top skills, 2020 | Top skills, 2015 |
|---|---|
| 1  Complex problem solving | 1  Complex problem solving |
| 2  Critical thinking | 2  Coordinating with others |
| 3  Creativity | 3  People management |
| 4  People management | 4  Critical thinking |
| 5  Coordinating with others | 5  Negotiation |
| 6  Emotional intelligence | 6  Quality control |
| 7  Judgment and decision-making | 7  Service orientation |
| 8  Service orientation | 8  Judgment and decision-making |
| 9  Negotiation | 9  Active listening |
| 10  Cognitive flexibility | 10  Creativity |

**Source:** WEF, 2018a.

Table 1.2 shows two lists of "top skills" identified in forecasts of this type. These lists are not accompanied by any information about the survey method used, or the sources, or the number of samples used to compile the forecast, or the occupations and activities for which these skills are supposed to be relevant.

Moreover, the skills themselves overlap in many cases, or are understood very differently depending on the context – industry or activity – in which they are applied. The skills listed here are too generic or general to be of use in matching; indeed, they are barely relevant for many occupations.

▶ **Figure 1.14.  Typical, current example from the skilling, upskilling and training area, showing further dimensions of skill definitions and their mapping/classification**

**Business Skills Collection**

Help employees develop the right hard and soft skills they need

Skillsoft's Business Skills collection contains Skillsoft's most rigorous and comprehensive content offerings, including state-of-the-art content such as critical and adaptive thinking, virtual collaboration, cross-cultural competency, and new media literacy. Combining that with industry-leading instructional technique and design and highly effective application resources, this content collection will drive their personal development while moving their organizations forward.

On the platform Percipio, the Business Skills collection contains over 65 curated channels that have been mapped to core competencies across popular topics such as:

▶ Business Operations
▶ Customer Service
▶ Professional Improvement
▶ Management & Leadership
▶ Project Management
▶ Sales & Marketing

**Source:** Author.

The example shown in figure 1.14 illustrates the typical, mostly generic and unspecific structuring and disordered classification of skills and soft skills into completely irrelevant categories such as "sales & marketing" or "project management".

Skills definitions, moreover, sometimes vary, or even have contradictory meanings, in different languages. There are also the difficulties attached to arriving at a reliable definition of levels: what does "good" or "very good" knowledge mean, and what distinguishes an "expert" in a certain skill? Is it theoretically acquired knowledge, for example, or is it knowledge already applied in a real professional environment? In contrast to other areas of big data, there are no binding scales and validations. As is clear from table 1.1 and figure 1.14, there is a pervasive pattern of conceptual ambiguity, lack of specificity and lack of concision.

NLP is the second decisive component, after knowledge representation, in the structuring and processing of data in the area of occupational big data. However, even though NLP has made great progress, and will continue to do so in the years ahead, the binding processing of language remains extremely difficult for the computer to manage. At the same time, other clear big data categories and criteria such as numbers, prices, GIS (geographic information systems) data and coordinates, temperatures, clicks and likes, products, elements of computer vision, pictures, etc. are not applicable here.

There follow here a few examples of "skills" descriptions from predictive evaluations and recommendations.

▶ **Digital skills:** What exactly does this mean? Does it include operating digital devices such as smartphones or computers, posting on Instagram, knowledge of complex building information modelling applications in real estate drafting and planning, or dealing with the Internet, Google Analytics, Facebook or social media? The definition of the skill is completely imprecise.

▶ **Marketing or digital marketing:** Here too, the term has thousands of different meanings and is far too broad and too unspecific. Therefore it is completely meaningless for almost all applications, especially for matching.

▶ **Project management skills:** This term too is almost completely useless. Almost everyone has project management skills on some level, but this knowledge cannot be compared or categorized. For example, the project management knowledge of a foreman on a large tunnel construction site, of a project manager of a large IT application, of a campaign manager in a public authority and of a process engineer or event manager are all very specific capabilities and thus obviously have very different meanings. It is nonsensical to comprise them all into a single "matchable" skill.

Skills are usually presented and processed without any relevant semantic context: and yet this context is the game-changer, which should go beyond what can be extracted linguistically and can be elucidated e.g. via text mining/NLP. Many applications are still based on keywords and strings, but not on context and multi-dimensional semantic concepts.

Moreover, in almost all current methods and standard classifications, the skills are often assigned to superordinate occupational groups. This, too, is a systemic error that urgently needs to be corrected. In the future, meaningful and helpful matchable skill sets should be defined and, if necessary, specifically promoted.

▶ **Figure 1.15  Typical relation of skills and professions in ESCO or O*NET**
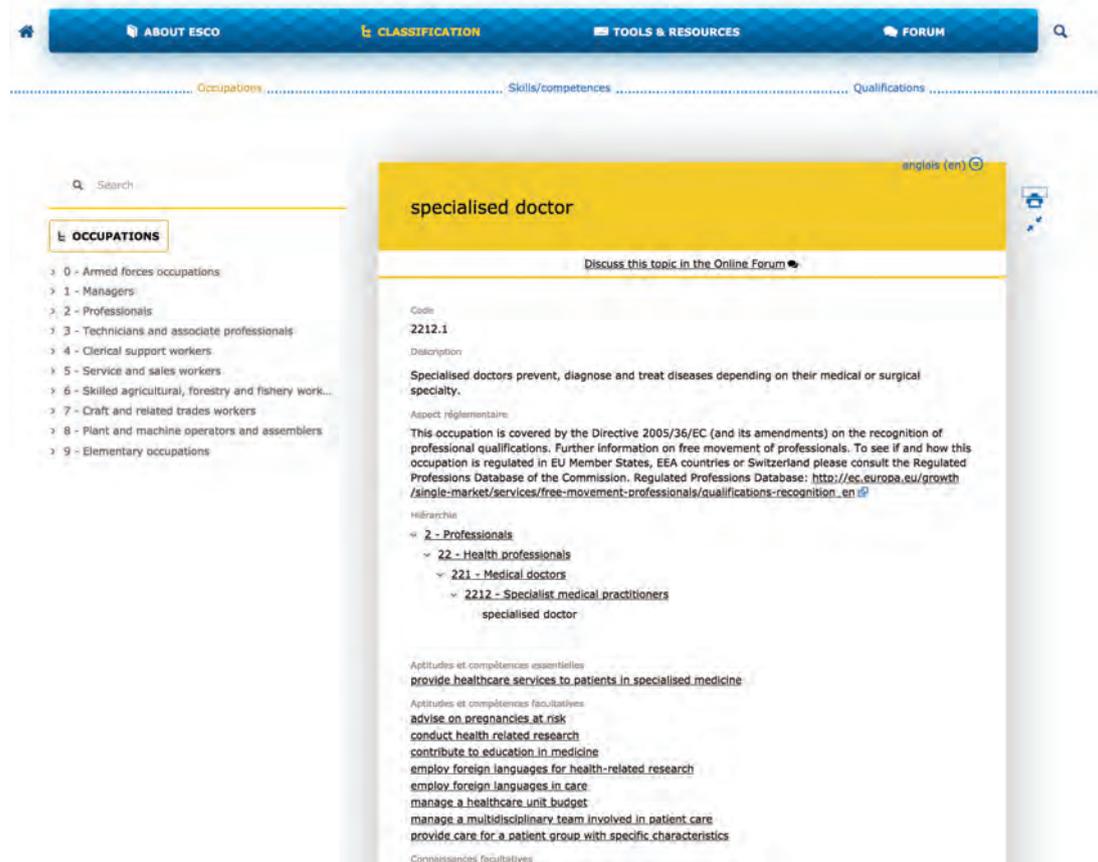


**Source:** Screenshot from ESCO (1st Oct. 2020).

The example of a medical specialist in figure 1.15 shows how irrelevant most facultative skills are for the various specialists belonging to the broader group. They are relevant only for specific doctors (in the case shown here, gynaecologists, surgeons or urologists).[16]

---

16  For a more detailed account of this problem, see: https://janzz.technology/ontology-and-taxonomy-stop-comparing-things-that-are-incomparable/ [16 Feb. 2020].

The questions we face in seeking to anticipate and match skills are many and complex. How many professions will disappear in the future, and which ones? What jobs will be affected by robotics and automation? And when? What share do these occupations have in the overall market? What are the relevant skills that will ensure the professional advancement of future generations? What are the most promising fields of study and vocational training for those seeking a secure future? Which skills and competencies should be part of the much-discussed "lifelong learning" and can they even be taught to a large number of people?

If we really want to use big data to help us answer these questions, as we seek to make progress in the matching of complex and multi-layered profiles and skills data, we must make sure that we methodically include the entire picture and as much context as possible.

## 1.4.4  Illustrative examples

Imagine a big data company that uses its technology to collect data of all available job advertisements from the English-speaking world over a period of ten years. It extracts all possible entities, above all, skills. The statistically most relevant advertisements are those from the United States, as they are the highest in quantity. Nevertheless, all geographical areas are modelled individually for separate evaluation.

Also relevant is the fact that advertisements from the United States are unique in their level of detail, which makes their comparison with other data sets almost impossible. Thus, when parsing advertisements from the United States one can find a lot of information set out explicitly that is available only implicitly or not at all in comparable data from the United Kingdom and other English-speaking regions.

A few examples (shown in figure 1.16) can illustrate specifically the quality and quantity challenge of extracting data from job specifications and advertisements.

▶ **Figure 1.16  Various examples of job advertisements**

---

### Curator–Botanist

*Royal Botanic Gardens Kew – Kew Gardens*
Apply On Company Site
Kew Gardens

£21,527–£23,282 a year

In addition to this, the role includes facilitating access for our many visitors to the collections from around the world, identifying African & Madagascan specimens and providing research support to the team.

Our successful candidate will have a botanical background with experience of working with Herbarium / natural history collections. You'll also need to have excellent interpersonal skills and enjoy both working in a team and independently.

The salary will be £21,527–£23,282 per annum, depending on skills and experience.

We offer a fantastic range of benefits including a broad range of Learning and Development opportunities, with access to the Civil Service training curriculum, generous annual leave entitlement for new starters, family friendly policies, a choice of competitive pensions and flexible benefits scheme.

If you are interested in this position, please submit your application through the online portal, by clicking "Apply for this job".

We are committed to equality of opportunity and welcome applications from all sections of the community. We guarantee to interview all disabled applicants who meet the essential criteria for the post.

No Agencies please.

---

# Carpenter

New York University
972 reviews – New York, NY
Apply On Company Site

## Position Summary

SHIFT: Monday–Friday 8:00AM–4:30PM

Primarily responsible for building, installing, repairing and refinishing building fixtures (e.g., doors, floors, stairs, ceilings, walls, windows, cabinets, etc.) of assorted materials (wood, ceramic tiles, concrete, stone, glass, plexiglass formica, metal, etc.).

## Qualifications

▶ *Required Education:*

High School diploma or equivalent

▶ *Preferred Education:*

Trade school or other post-elementary education.

▶ *Required Experience:*

Three years of hands-on carpentry work experience for general building maintenance.

▶ *Preferred Experience:*

Experience in higher education, health care, telecom or other sensitive facilities where high levels of performance and reliability are required.

▶ *Required Skills, Knowledge and Abilities:*

The ability to use all power and hand tools. The ability to read mechanical sketches and shop drawings. Basic arithmetical and measuring skills. Common knowledge of materials. Good communication skills with the ability to follow verbal and written instructions. Positive attitude, willingness to collaborate and team with others to accomplish work efficiently and to high quality standards. Proficient in English (e.g., reading directions and instructions written in English). The ability to lift and carry up to 50 pounds, climb a ladder, work in small, cramped spaces, and in outdoor environments and elements (combined 40%–50%). MUST BE AVAILABLE TO WORK NIGHTS AND/OR WEEKENDS.

▶ *Preferred Skills, Knowledge and Abilities:*

Knowledge of University and department policies, processes, and quality standards.

▶ *Additional Information*

EOE/AA/Minorities/Females/Vet/Disabled/Sexual Orientation/Gender Identity

<div style="border:1px solid">

## Möbelschreiner / Monteur (m/w)

Job3000 – Zürich, ZH
Weiter zur Bewerbung
Zürich, ZH

**Ihr Profil:**
- Ausbildung als Schreiner EFZ oder erfahrener Möbelmonteur
- Umzugserfahrung
- Körperlich top fit
- Gute Deutschkenntnisse von Vorteil
- Führerausweis Kat. B, B/E von grossem Vorteil
- Flexibel und belastbar

**Ihre Aufgaben sind:**
- Transport und Montage von teuren Designermöbel
- Ein-und Ausladen
- Montagearbeiten
- Mithilfe in der Entsorgung
- Ein- und Auspacken

Fühlen Sie sich angesprochen? Wir freuen uns auf Ihre vollständigen Bewerbungsunterlagen (Lebenslauf mit Foto, Zeugnisse, Diplome), welche sie an die folgende E-Mail Adresse schicken können:

</div>

<div style="border:1px solid">

## Carpintero autónomo

**AVERALIA – Madrid, Madrid provincia**

Solicitar
Madrid, Madrid provincia

24.000€–36.000€ al año

Se necesitan operarios autónomos
con amplia disponibilidad
y que pueda hacer una o dos guardias
al mes los fines de semana.

Somos una empresa que trabaja
para asistencias de seguros.

Tipo de puesto: Jornada completa

Salario: 24.000,00€ a 36.000,00€ /año

</div>

**Source:** job advertisements published on indeed.com.

## The challenge of these examples

It is abundantly clear from even these few simple examples that the density and relevance of the information given vary massively. The examples also show that much of the information (approximately 30–40 per cent) is often implicit, contained in stipulations about education/training, qualifications and experience. If such information is not represented in an accurate and semantically meaningful manner, the collected data is quickly distorted: pseudo-correlations and statistical significances seem to arise when in fact they do not exist at all. In addition, the following deviations are to be expected when using such a survey method:

▶ approximately 30 per cent of jobs are assigned informally, anonymously and without the use of a job advertisement;

▶ large companies or those with a duty to publish vacancies are statistically over-represented;

▶ technical/ICT-related occupations provide more specific skills, as in this domain it is more common to mention all the necessary tools and technologies;

▶ segments and industries that suffer from a shortage of skilled workers tend to advertise jobs on several platforms; at the same time, a large number of advertisements are not aimed at hiring people with the skills sought, but are placed only for employer branding or self-marketing purposes, and yet are hardly distinguishable from real job advertisements;

▶ implicit information is recorded only to a limited extent and is statistically under-represented, despite its high relevance;

▶ the extracted data is often very imprecise and incomplete; job titles cannot be assigned or mapped correctly due to cryptic formulations.

Combining this body of data with other assumptions and criteria, such as assumptions about the forthcoming automation of different professions and sectors (where a note of caution needs to be struck about the relationship between correlation and causality), results in a forecast with a very high probability of error, based on data that are hardly suitable for matching anyway.

This kind of process results in decision-making and orientation aids, which are then conveyed in the form of summaries and lists, such as that shown in box 1.1, for example. Can these really say anything about the future of skills and abilities? Take, for example, the skill "mental health": is this a single skill at all? Do the skills not lie, rather, in e.g. treatment of mental health, mental health care, etc.? How could "mental health" be suitable for matching profiles and multi-dimensional data?

---

▶ **Box 1.1: Which ten skills are now mentioned more frequently?**

These skills have shown the fastest growth in the number of mentions between 2012–14 and 2014–16. Several factors can drive growth, including an increase in the number of vacancies for the job that requires this skill, or an increase in the range of jobs that use this skill.

A number of these skills (looking at the top 100) relate to caring for others, such as patient care, mental health, and working with patients who have dementia. A second group of skills reflect the opportunities and threats that come from living in a more connected world. These include digital marketing, big data, social media, information security and firewalls.

1. Big data
2. Information technology industry experience
3. Contract accountancy
4. Onboarding
5. Digital marketing
6. Information security
7. Transportation logistics
8. Front-end development
9. Patient care
10. Mental health

**Source:** Author.

---

## 1.4.5. Conclusion

### The collection of skills and other occupational data from big data does not result in the creation of high-quality smart data

We are convinced that a purely big-data-based prognosis of future skills will hardly be possible in a way that goes beyond the obvious. These forecasts will be just as unspecific and mostly just as wrong as current long-term weather forecasts. If, however, the quality and consistency of the data should improve significantly in the years ahead, and at the same time the extremely high variability can be reduced, we can certainly expect somewhat more reliable and meaningful predictions. The methodology used to collect and structure data plays a decisive role here.

## Implicit information is at least as important as explicit, readily available information

The discussion above has highlighted the enormous variance in information density and relevance across vacancy postings. Much of the information about the job advertised is often implicit, hidden in stipulations about education/training and qualifications, and about experience. If these are not represented accurately, semantically and in a knowledge representation, the collected data is distorted. Importantly, geographical differences in the way jobs are described, and the level of detail included, make comparison with other data sets almost impossible.

If we want to make progress in the assessment of skills relevance, in connection with new hires or the employability of candidates, it will be necessary to consistently check and record which criteria and skills were decisive and why a candidate was finally hired. Were these criteria and skills congruent with the specifications of the job advertisement? Were they different, because the person hired already possessed certain skills required for the new position, so that complementary criteria weighed in the decision? Are these criteria reviewed periodically? Which factors are responsible for good or bad performance? Which skills seemed important at the time of recruitment or placement, but were irrelevant for the actual job performance? What are the real skills gaps? Have there been significant changes in the framework over time that will influence decisions on appointments or assessment of similar jobs in the future? All these criteria would have to be gathered in a data collection and evaluated in order to facilitate the development of significant and precise big data models and AI processes.

## How skill sets can be accurately modelled for both emerging and advanced employment markets

With regard to matching, an acceptable quality level with improved results can be achieved relatively quickly by means of a close integration of knowledge representation with big data applications. However, if attempted with big data and machine-learning/AI approaches that are lacking in prior standardization, systematic structuring and semantic expansion, this undertaking will not succeed.

All this explains why it will be mainly experts – and not algorithms – who will continue to be responsible for such data modelling in the future.

**2**

# Using big data to assess and meet skills needs: Learning from advanced countries' experience

This chapter collects insights and experiences on the actual use of big data for the analysis of skills requirements in the labour market in advanced economies. It provides some useful insights, drawing on academic research as well as practical applications by various institutions.
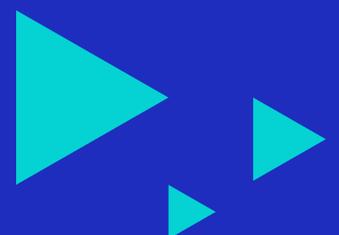
## ▶ 2.1. Using big data and AI for identifying LMI in Austria[17]

**Claudia Plaimauer**

### 2.1.1. Preliminary experience with big data and AI methods

As a long-term contractor of the Austrian public employment service (PES), 3s has – inter alia – developed online databases on general vocational information,[18] and on developments in labour market requirements,[19] as well as elaborate taxonomies for organizing both occupations and skills and competencies. In developing these systems, we relied mostly on traditional methods and data sources.

However, in 2013, on discovering Jobfeed,[20] Textkernel's big data platform for comprehensively investigating national online vacancy markets, we became determined not only to exploit automatically processed online vacancy data to analyse labour market demand, but also to test the potential of AI methods in several other contexts. Our preliminary experience covers activities such as the following:

▶ **Automated normalization of free text survey results:** For many years 3s had analysed the Austrian PES' biannual enterprise survey.[21] In the part of this work that involved normalizing registered occupations, skills, competencies and training needs, we had relied on an automation-supported procedure (text string matching). In 2013 we tested Textkernel's natural language processing (NLP) approach as an alternative, and this turned out to produce far better results.

▶ **Capturing Austria's online job market:** In 2015, 3s cooperated with Textkernel in setting up a first Austrian version of Jobfeed.[22] The Austrian PES, as a key customer, supported its development and was rewarded with tailor-made query options, e.g. the possibility of searching and analysing the data on the basis of the PES' proprietary labour market taxonomies.[23]

▶ **Comparison of VET qualifications:** In 2018 and 2019, 3s contributed to European research projects investigating the extent to which AI methods could facilitate a translingual comparison of European VET qualifications.[24]

▶ **Curriculum development:** 3s increasingly uses Jobfeed-based analyses as additional input when counselling curriculum developers at universities, including universities of applied science.

▶ **Testing NLP performance:** In 2014–15, in the context of Cedefop's mid-term skills supply and demand forecasts, 3s conducted a parsing test with several commercial providers, generating insight into the comprehensibility and accuracy achieved by semantic technologies when processing job titles and requirement specifications from the text of OJVs.[25]

▶ **Online vacancy analysis:** In 2017 the Austrian Jobfeed vacancy database was comprehensive enough to serve as a data source for the PES' annual online vacancy analysis. Since then 3s has annually evaluated data exports from this source, representing demand information at the level of occupations and

---

17 This section is based on preliminary experience at 3s Unternehmensberatung GmbH and the Austrian Public Employment Service.

18 http://bis.ams.or.at [13 Feb. 2020].

19 AMS Skills Barometer, http://bis.ams.or.at/qualibarometer [13 Feb. 2020].

20 https://us.jobfeed.com/home.php?country_selection [13 Feb. 2020].

21 For further information see Ziegler and Auzinger, 2012.

22 www.jobfeed.com/at/home.php [13 Feb. 2020].

23 For further information, see Plaimauer, 2016.

24 For further information, see Auzinger et al., forthcoming.

25 For further information, see Plaimauer, 2015.

federal provinces, while also distinguishing between vacancies posted at the Austrian PES' eJob-Room[26] and those not (also) posted there. Results are reported in a format which can be fully integrated into the PES' "skills barometer".[27]

▶ **Testing AI methods for developing labour market taxonomies:** In 2017, 3s conducted a pilot project testing AI methods for validating and supplementing labour market taxonomies – an initiative discussed in more detail in the next section.

## 2.1.2. Using big data analysis to develop labour market taxonomies: The case of the Austrian PES' skills taxonomy

The Austrian PES uses two distinct taxonomies:

▶ an occupational classification (AMS-Berufssystematik[28]); and

▶ a compilation of occupational requirements (AMS-Kompetenzenklassifikation[29]), organizing skills, competencies, abilities, knowledge of tools or materials, specific work experience, attitudes, values, certificates, etc.[30]

Both taxonomies were designed by 3s around the turn of the millennium and have been maintained by 3s continuously since then. Figure 2.1 summarizes the main characteristics of these two systems.

▶ **Figure 2.1. Main characteristics of the Austrian PES' two labour market taxonomies**



**Occupational taxonomy**

▶ Occupations
▶ 1999/2000-
▶ 13,500+ concepts
▶ 84,000+ terms

**"Skills" taxonomy**

▶ Occupational requirements
▶ 2000/2001-
▶ 17,500+ concepts
▶ 29,000+ terms

▶ Goal: Comprehensiveness, high actuality, clarity, descriptiveness, uniformity;
▶ proximity to everyday language;
▶ Structure: Thesaurus and taxonomy;
▶ Usage context: Labour market information / matching / research
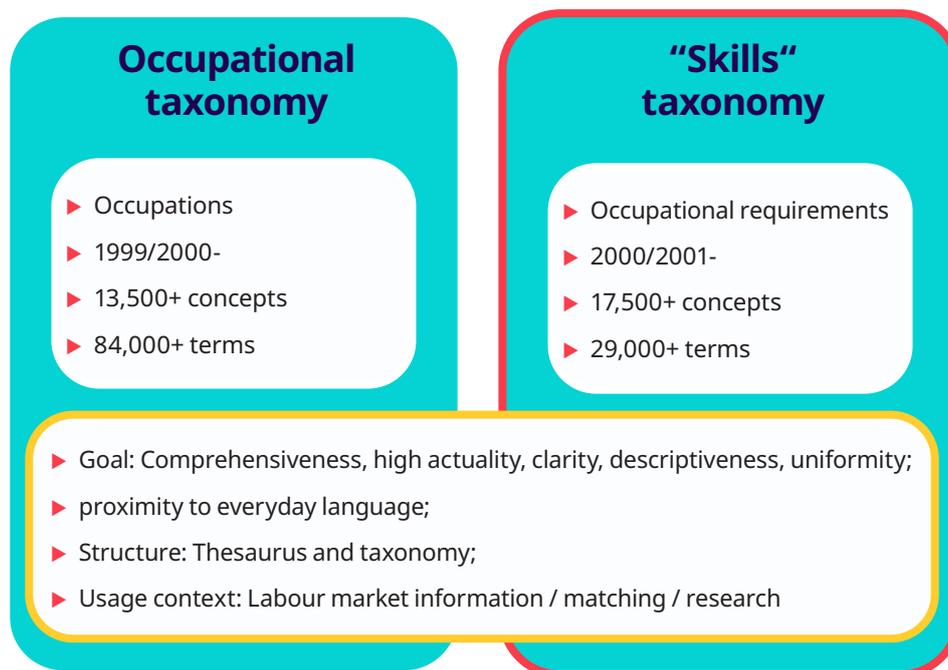
**Source:** Author.

---

26 This job platform contains not only all vacancies recorded in the PES' placement applications but also vacancies posted there autonomously by employers: see https://jobroom.ams.or.at/jobroom/login_as.jsp [13 Feb. 2020].

27 For further information, see Plaimauer, 2018a, b.

28 Available online at: https://www.ams.at/bis/bis/BerufsstrukturBaum.php [13 Feb. 2020].

29 Available online at: https://www.ams.at/bis/bis/KompetenzstrukturBaum.php [13 Feb. 2020].

30 For the sake of brevity we subsequently refer to this system as the "skills" taxonomy – the quotation marks signalling that this taxonomy contains far more than just skills.

The two taxonomies are well established, comprehensive, and interlinked to form occupational skills profiles used for the purposes of LMI, research and matching people to jobs. The vocabulary is organized as a thesaurus, supplemented by an additional classificatory structure; this structure is guided by working rules, but also to an extent by practical needs arising from the contexts in which they are applied. Both taxonomies offer the abundance of detail that is needed to characterize occupations in a descriptive manner; they also provide several layers of more generic concepts to summarize and structure this detail, e.g. for display purposes or to enable systematic searches. Figure 2.2 gives an example of how the organization into broader and narrower thesaurus concepts is used to generate a structured display of occupational requirements.

▶ **Figure 2.2. Presentation of "skills" in occupational profiles: An example**



**Boat and ship builder**

Required transversal competencies
- ▶ dexterity
- ▶ physical resilience

Required specialist skills
- ▶ use manufacturing plants, especially
  - ▶ machine tools
  - ▶ polymer processing machines
- ▶ material processing skills, especially
  - ▶ manual processing techniques
  - ▶ grinding technology
- ▶ woodworking skills, especially
  - ▶ slicing
  - ▶ gluing
  - ▶ ...

Occupation

Occupational requirements

**Source:** Author.

In maintaining the occupational and "skills" taxonomies, 3s relied until recently on methods traditionally applied in this context, such as:

- ▶ functional analysis of occupations and occupational tasks;
- ▶ comparative assessment of taxonomies of related content;
- ▶ semantic analysis of terms and their relations;
- ▶ terminology checks against taxonomy guidelines ("vocabulary control");
- ▶ editorial evaluation of user feedback;
- ▶ analysis of curricula;
- ▶ editorial evaluation of vacancy texts;
- ▶ evaluation of transaction logs.

In 2017 this portfolio of tools was extended to include AI methods: specifically, 3s tasked Textkernel with producing automatically generated input for the maintenance of the "skills" taxonomy, addressing in particular the following goals and research questions.

▶ **Goal 1: Validating taxonomy terms**

  ▶ Are the taxonomy terms chosen to represent occupational requirements ("skills") actually used in vacancies?

  ▶ Which linguistic formats occur frequently in vacancies, and which rarely, or never?

▶ **Goal 2: Supplementing the "skills" taxonomy**

  ▶ Are any "skills" terms commonly used in vacancies missing?

  ▶ Is there any indication that "skills" concepts are missing?

▶ **Goal 3: Gaining insight into the taxonomy's suitability for NLP**

  ▶ Does the current format of the taxonomy in any way impede its use in NLP?

Below we describe the methods applied for addressing these research goals, present the insights achieved and interpret their relevance for the further development of the "skills" taxonomy.

## Goal 1: Validating taxonomy terms

### The language of the "skills" taxonomy differs from the language used by recruiters

Textkernel applied text string matching, a fully automated procedure, to compare the complete vocabulary of the Austrian "skills" taxonomy (over 29,000 terms) with the requirements set out in the texts of all Austrian online vacancies published in the preceding year (over 850,000 unique vacancies), recording the frequency of occurrence of every single taxonomy term. The results provided interesting insights into the difference between the language used by recruiters and the controlled vocabulary of our "skills" taxonomy. For example, *56 per cent of the Austrian "skills" terms did not appear at all in the job ads of the preceding year*.

Does this mean that our "skills" taxonomy is poorly aligned with the language used by recruiters? We think not, for several reasons. First of all, frequency of occurrence in vacancies is only a rough proxy for a skill's labour market relevance, considering that job advertisements rarely if ever contain a comprehensive description of employers' requirements. Furthermore, the intention of our "skills" taxonomy is not to duplicate but to interpret and structure the reality of the labour market; it aims at building a comprehensive model of this specific domain, and thus unavoidably also contains concepts that have no (or only little) observable labour market visibility,[31] such as:

▶ skills required in occupations for which recruitment is rarely done via online vacancies (e.g. religious professionals) or which have not been in demand during the preceding year;

▶ skills tacitly expected from applicants: e.g. *Holzfräsen* (timber milling) on the part of joiners;

▶ generic taxonomy concepts needed for summarizing detail: e.g. *Bürosoftware-Anwendungskenntnisse* (basic office software skills) as a broader term for skills such as "word processing", "table calculation" or "presentation software";

▶ explicit reformulations of otherwise ambiguous terms, e.g. "MAP (Mobile Access Protocol)" to be distinguished from "MAP (Mobile Application Part)".

Also, the Austrian "skills" taxonomy is a controlled vocabulary, ensuring that all terms it contains have a unique reference, are phrased in a uniform format, are balanced with respect to gender and are free from discriminatory language, whereas naturally occurring language is often vague, biased and stylistically varied.

---

31  Or at least no visibility within a twelve-month observation period.

Therefore, a certain degree of deviation between the two language uses is not only to be expected, but also desirable.

Nevertheless, if we want to ensure that the Austrian "skills" taxonomy is also suitable for automatically processing natural language (see Goal 3), e.g. in automated matching of CVs and job advertisements, then the language of the vacancy market should also be represented as far as possible. Under "Goal 2: Supplementing the 'skills' taxonomy" below, we will describe which of these terms were found to be missing.

### The most frequently occurring taxonomy terms tend to be short and transversal (or at least cross-sectoral)

Only 2.6 per cent of the overall "skills" vocabulary appeared in more than 1,000 vacancies during the preceding year. The ten most frequently occurring taxonomy terms tend to be short and transversal. They were, in decreasing order of frequency:

▶ *Berufserfahrung* (occupational experience);

▶ *Erfahrung* (experience);

▶ *Deutschkenntnisse* (knowledge of German);

▶ *Service* (customer service);

▶ *Praxis* (experience);

▶ MS Office;

▶ HTTP;

▶ *Englisch* (English);

▶ *Reisebereitschaft* (willingness to travel);

▶ *Belastbarkeit* (resilience).

### Term length and frequency of occurrence are negatively correlated

The more characters were needed to express a taxonomy term, the less frequently it appeared in vacancies (see figure 2.3). This finding is a result of the investigation done for the Austrian PES, using their "Skills" taxonomy and vacancies for jobs in Austria, but it also has been observed for other languages (e.g. English) and other contexts as well.

Even so, there are a few taxonomy terms containing more than 25 characters which nevertheless appeared often in job advertisements (see table 2.1). These almost exclusively represent either transversal/cross-sectoral skills or generically expressed formal qualifications, and they tend to be adjective–noun (e.g. *betriebswirtschaftliche Kenntnisse*) or (adjective–)noun–noun combinations (e.g. *Unternehmerisches Denken und Handeln, Steuerungs- und Regelungstechnik*), or simple prepositional phrases (e.g. *Freude am Umgang mit Menschen*).

Our conclusion for future taxonomy management is to strive for shorter, less contextualized concept names; where longer and more contextual terms cannot be avoided, they should at least be supplemented with brief alternative names, and/or subordinated to a shorter more generic concept: for example, the concept name *Schweißen von zweiseitigen Stumpfnähten* ("welding of double-sided blunt seams") would be linked to the broader concept *Schweißen* ("welding").

### Some linguistic forms never appeared in advertised vacancies

Terms containing components in parentheses were never used in job advertisements. These include such forms as:

▶ terms with explanatory adjuncts, e.g. "courier (mail transfer agent)";

▶ terms with alternative abbreviated names, e.g. "computer based training (CBT)";

▶ **Figure 2.3. Correlation between number of characters in term and number of occurrences (excluding 0 occurrences)**
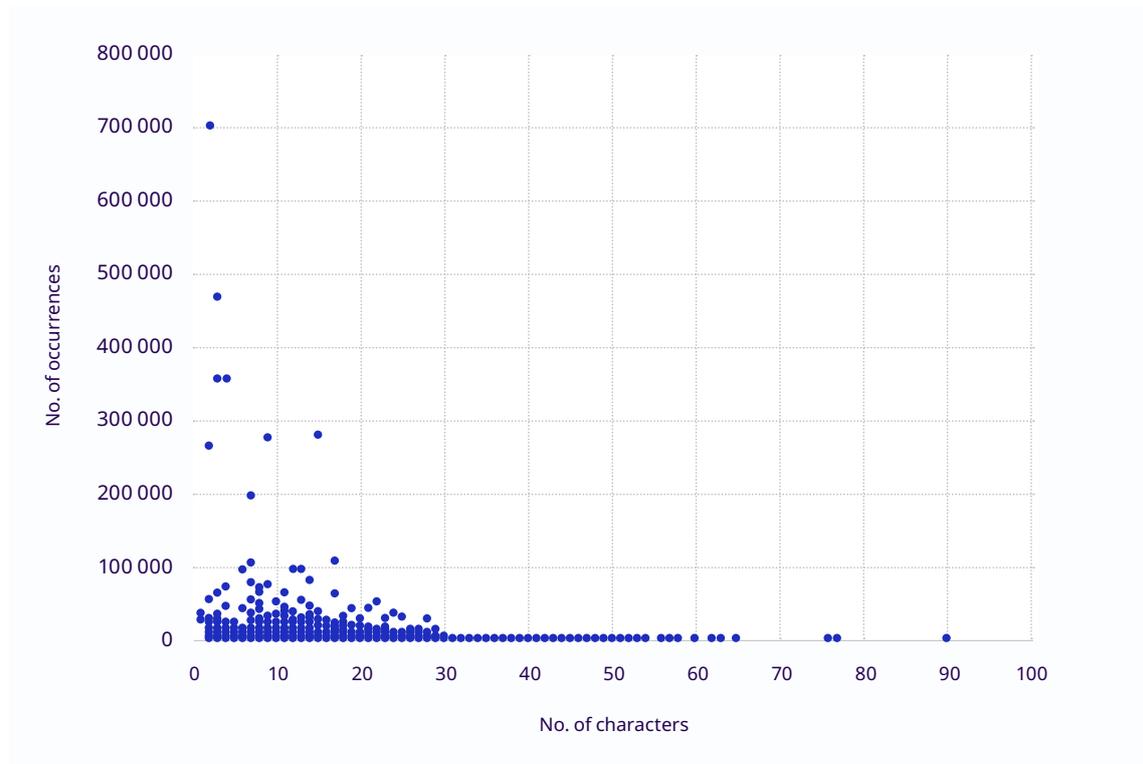


**Source:** Textkernel.

▶ **Table 2.1. Longer but frequently occurring taxonomy terms**

| Taxonomy term (English translation) | No. of characters | No. of occurrences |
| --- | --- | --- |
| Unternehmerisches Denken und Handeln (entrepreneurial competences) | 36 | 1 808 |
| Abgeschlossenes technisches Studium (completed technical studies) | 35 | 1 034 |
| Betriebswirtschaftliche Kenntnisse (business skills) | 34 | 1 066 |
| Abgeschlossenes Wirtschaftsstudium (completed economic studies) | 34 | 882 |
| Steuerungs- und Regelungstechnik (monitoring and control technology) | 32 | 653 |
| Serviceorientierte Arbeitsweise (service-oriented method of working) | 31 | 475 |
| Betriebswirtschaftliches Denken (business thinking) | 31 | 440 |
| Prozessorientierte Arbeitsweise (process-oriented way of working) | 31 | 436 |
| Interesse an neuen Technologien (interest in new technologies) | 31 | 423 |
| Freude am Umgang mit Menschen (pleasure in dealing with people) | 29 | 13 863 |
| Englisch in Wort und Schrift (command of spoken and written English) | 28 | 10 739 |
| Strukturierte Arbeitsweise (structured work style) | 26 | 11 652 |

**Source:** Author.

▶ terms specifying the skill level, e.g. Brillengläser vermessen (Grundkenntnisse) ("measuring dioptric strength of glasses (basic skills)");

▶ formally disambiguated terms, e.g. "ASP (active server pages)", "ASP (application service provider)".

Since all but the last example given here could easily be presented using other terminological strategies, we recommend that add-ons in parentheses should be used solely for disambiguating homographs.

## Goal 2: Supplementing the "skills" taxonomy

To generate candidates for addition to the "skills" taxonomy, Textkernel applied a semi-automated proce-dure. First, all vacancies registered in Jobfeed Austria during the preceding twelve months were searched for skills, using keyword extraction and an identification method based on a set of skill-denoting linguistic patterns appropriate for vacancy texts. This process was enhanced by a cross-validation approach that looked for further evidence that a given text string denotes a skill in different sources, CVs or Textkernel's proprietary ontology, which had been compiled bottom-up from millions of job vacancies. This evidence aggregation was performed in a semi-automatic way using both machine-learning models and human quality assessment.

The next processing step involved data cleaning. Spelling variants were detected to identify terms with the same meaning but slightly different wording, both in the newly mined skills and in the Austrian "skills" taxonomy. Also, strings that formed part of complete phrases were taken into account, because these could indicate either broader/narrower term relations or spelling variants. Spelling variants were excluded from the output to guarantee a high number of truly new "skills".

The newly mined "skills" were then run on one year of vacancy data from Jobfeed Austria. The occurrence of each skill per vacancy was counted; terms occurring more frequently than twelve times[32] were com-pared against the Austrian "skills" taxonomy, and if they did not yet appear there, were added to a list of potential candidates for inclusion. This list of potential new taxonomy terms was then checked manually to detect non-skill content, e.g. terms that occurred very frequently and had not yet been discarded, such as Arbeitsgenehmigung ("work permit"), terms that are so general to be unsuitable for inclusion in the "skills" taxonomy, such as Arbeitswelt ("world of work") or terms that represented only insignificantly deviating variants of already detected new "skills". In addition, the body of all detected "skills" terms was adjusted by applying part-of-speech tagging.

In the final step of the process, placement suggestions within the Austrian "skills" taxonomy were generated for the newly mined terms, to facilitate subsequent editorial processing. Two approaches were applied here:

▶ at string level, the new terms were compared to the Austrian "skills" taxonomy;

▶ at a more semantic level, the new terms were compared to Textkernel's ontology, exploiting the me-ta-information contained therein.

Terms for which no placement suggestion at all could be generated were marked as most likely candidates for new "skills" concepts.

On top of placement suggestions, the final list of approximately 1,900 terms also contained:

▶ the most frequently detected occupational context for each term;

▶ a rough classification into types of "skill" term, distinguishing between "soft skill", "technical skill", "broader term", "new skill" and "synonym".

The outcome of this procedure provided us with the following insights:

### AI generated valuable evidence-based amendment potential

Although considerable additional human processing was necessary to review and supplement the taxon-omy, and then integrate new terms or concepts into it, automated skills mining not only generated several hundred terms previously missing from the AMS Kompetenzenklassifikation, it also provided an evidence base for taxonomy maintenance decisions, which up to then had tended to lean somewhat towards the subjective.

---

32  This threshold was chosen on practical grounds: "skills" occurring less frequently tended to contain more noise (e.g. spelling variants, morphological variants).

## The Austrian "skills" taxonomy is fairly comprehensive

The relatively small number of approximately 1,900 potentially missing "skills" terms suggests that the Austrian taxonomy is already fairly comprehensive. Of the candidates for addition detected, about 900 terms resembled specialist "skills"; of these, 635 were selected to be added as naming alternatives to existing "skills" concepts; another 97 terms were confirmed as truly missing concepts and thus added to the taxonomy.

Most of the new "skills" concepts were highly specific, e.g.

▶ "Poka yoke" (a quality assurance technique);

▶ "Haproxy" (open-source software offering high availability, load balancing, and proxying for TCP and HTTP-based applications);

▶ "webtracking" (collecting user information).

▶ Supplementary alternative names for already existing "skills" concepts included:

▶ Lebensmitteltechnik ("food technology") as alternative to the existing Lebensmitteltechnologie;

▶ "industrial engineering" as alternative to the existing Fertigungstechnik (manufacturing technology);

▶ "cloud services" as alternative to the already existing "cloud computing".

It would not have been feasible to produce such highly specific additions without AI support.

## Skills mining is effective, but it comes at a price

Text mining is a highly effective method for identifying evidence-based amendment needs for taxonomies, but it comes at a price: the technical procedure and the subsequent post-processing by human taxonomy editors was so costly that it will probably be repeated only at long intervals.

## The Austrian "skills" taxonomy could be better customized for NLP

Skills mining provided further evidence that the Austrian "skills" taxonomy could be improved to render it better suited to the automated processing of vacancy text. Also, the observed negative correlation between term length and frequency of occurrence in candidates for addition underlined the need to strive for brevity and concision in taxonomy terms. With respect to linguistic form, candidate additions tended to have the format of simple or compound nouns (e.g. *Führungsverantwortung*)[33], of adjective–noun combinations (e.g. *durchsetzungsstarke Persönlichkeit*)[34] or of simple prepositional phrases (e.g. *Spaß am Verkauf*)[35]. Frequently occurring candidates almost all showed these formats – indicating that these linguistic forms should be preferred by a taxonomy customized also for NLP.

---

33 "Führungsverantwortung" – German for "management responsibility".

34 "durchsetzungsstarke Persönlichkeit" – German for "assertive personality".

35 "Spaß am Verkauf" – German for "Pleasure in selling".

# ▶ 2.2. The use of big data for skills anticipation and matching in the Netherlands[36]

**Renier van Gelooven**

## 2.2.1. Introduction to SBB

SBB is the Dutch Foundation for Cooperation on Vocational Education, Training and the Labour Market. Within the Foundation, vocational education and training (VET) and the labour market cooperate at national, sectoral and regional level.

The objectives of SBB are:

▶ to provide students with the best possible practical training with a view to their prospect of employment;

▶ to ensure that companies can employ the professionals they need, now and in the future.

According to its legislative mandate under the Dutch Act on Adult and Vocational Education, SBB works together with VET schools and the labour market in performing the following tasks:

▶ advising, accrediting and coaching work placement companies;

▶ developing and maintaining the qualification structure;

▶ providing research and information on the labour market, on work placement and on the efficiency of VET programmes.

SBB also advises the government on linking vocational education with the labour market.

### The SBB research and information team

The objective of our research and information team is to provide all relevant stakeholders with authoritative, recent, objective and relevant information about the fit between education and the labour market, VET and enterprise efficiency. We do this not only for, but together with these stakeholders. Our particular focus is on middle-level applied education. We provide information on both national and regional levels; about specific qualifications; and also at sectoral level. Above all, the information we provide is future-oriented.

## 2.2.2. LMI and the information pyramid

This section explains how SBB uses the information pyramid (figure 2.4), loosely based on Rodenberg's model for marketing intelligence (Rodenberg, 2002).

This pyramid contains several levels and explicitly takes into account the transfer of material between these levels. At the top of the pyramid we find "Action", indicating that our work should lead to things being done. These actions are based on rational, preferably data-driven, policies or advice, which in turn are based on knowledge. That underlying knowledge is drawn from information, which is created through combining a variety of data sources. Information is based on quantitative and also qualitative data, and the data originate from both internal and external sources.

---

36  This section is based on the experience of SBB, specifically its LMI programme.

▶ **Figure 2.4. The information pyramid**
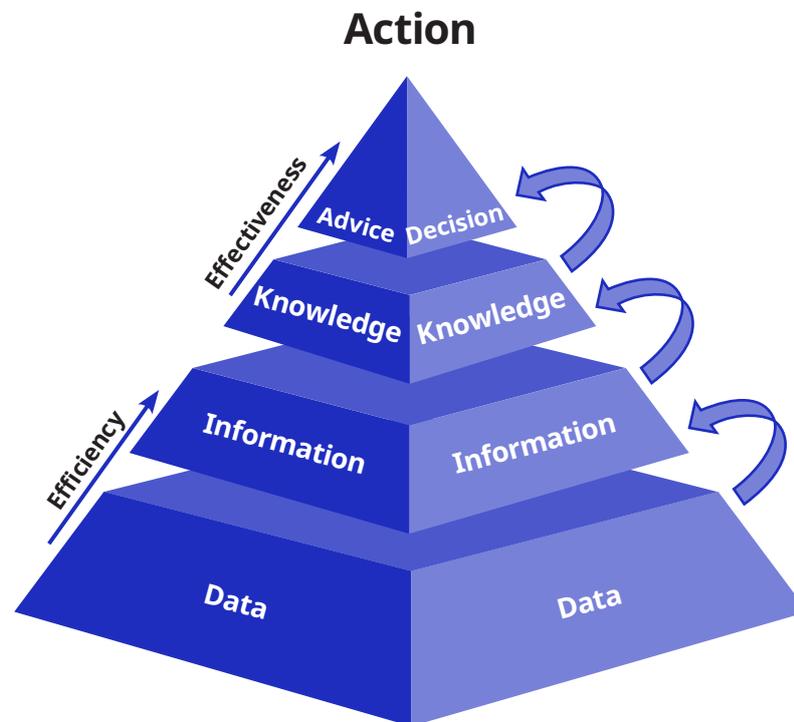


**Source:** Rodenberg, 2002.

## An intelligent network

To ensure a successful transfer of information to knowledge, and for that knowledge actually to have an impact, we need more than data. We therefore try to create an intelligent network using LMI. This entails discerning relevant patterns, identifying changes in these patterns, and being able to anticipate developments successfully. To do this, we need theory, context and commitment:

▶ theory on the allocation mechanisms that function between education and the labour market (Karsten, 2016, p. 124);

▶ contextual information about the specific situations and stakeholders involved, including their different interests;

▶ commitment on the part of those stakeholders (schools, trainers, companies, governments, etc.) to really make an effort and a contribution.

Big data can contribute, but only if aligned with this intelligent network. To bring about this alignment, SBB has founded online communities involving our main stakeholders, experts and our own work placement advisers. These "share your knowledge communities" are an essential part of our current research projects.

## SBB's data sources

In our work we use and combine a variety of data sources. The main sources that we use are:

▶ data from the educational system (e.g. numbers and attributes of students, graduates, etc.);

▶ data from the Central Bureau of Statistics (demographic, labour market, economic, etc.);

▶ data from the Employee Insurance Agency on (trends in) unemployment, job-change patterns, labour market outflow through (mainly) retirement (demographic trends);

▶ data from vacancies (Internet spiders).

We are constantly looking for other sources that may be useful and suitable.

## SBB's network

Besides these data sources, we have unique access to data and insights from or through:

▶ SBB's work placement advisers (450, distributed across regions and sectors);

▶ work placement companies and trainers (250,000, distributed across all regions and sectors);

▶ VET schools (70 institutions);

▶ students (500,000);

▶ experts.

## Big data

Big data – defined by Gartner[37] in three key words, volume, velocity and variety – is attracting increasing attention, as indicated by the number of searches on the Internet (figure 2.5, blue line).

▶ **Figure 2.5.  Numbers of Internet searches for "big data" and "education", 2008–15**



interest in the course of time ?

Average   1 jan. 2008   1 nov. 2011   1 sep. 2015

▶ Big data   ▶ Education
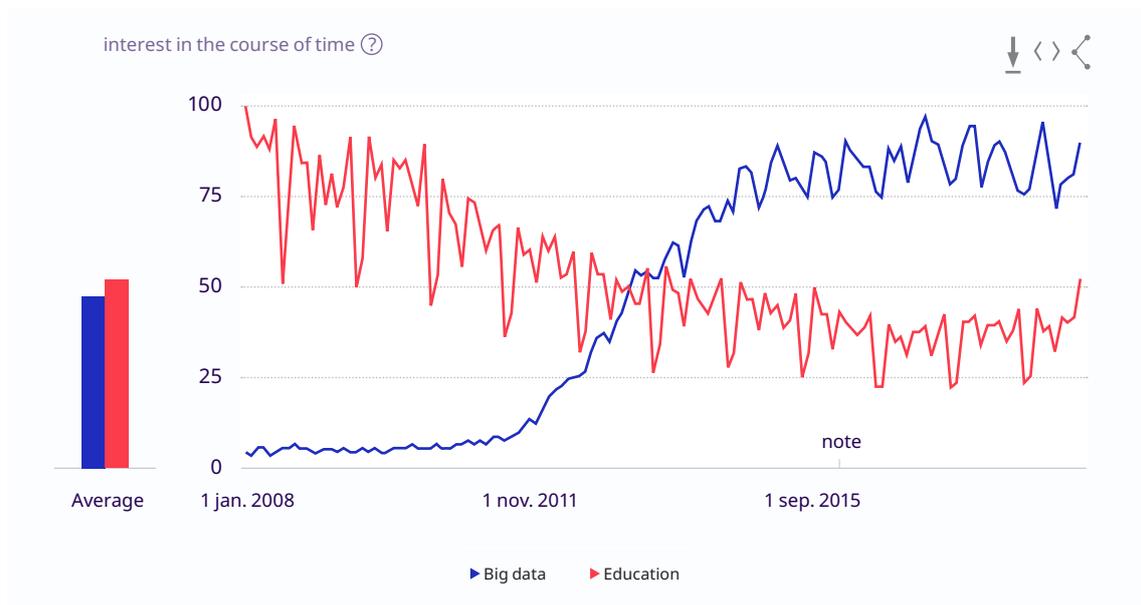
**Source:** Author.

In the field of skills and LMI data is copious and varied: that is, volume and variety are high. The velocity of data in this area, however, is not very high. On the other hand, the horizon for useful predictions is usually long-term, and the impact of actions is usually apparent only after the passage of years.

The Netherlands Ministry of Education explored the theme of big data some time ago (Bongers, Jager and te Velde, 2015). Their findings, summarized in terms of opportunities and threats, are shown in table 2.2.

---

37  https://www.gartner.com/en/information-technology/glossary/big-data [14 Aug. 2020].

▶ **Table 2.2.  Opportunities and threats in the potential use of big data, as seen by the Netherlands Ministry of Education**

| Opportunities | Threats |
| --- | --- |
| Better predictions | Privacy issues |
| Realization of innovations | Lack of data scientists |
| Higher productivity | Information overload |
| More evidence-based policies | Does not give a complete picture |
| New methods and sources for research | Risk of bad science |

**Source:** Bongers, Jager and te Velde, 2015.

The use of big data as part of LMI is still in its infancy. In the Netherlands, SBB is participating in several projects in this field. The Central Bureau of Statistics is researching the possibilities of using vacancy data and machine learning to detect patterns in emerging and disappearing skills (the "Werkinzicht" project). TNO (the Netherlands Organization for Applied Scientific Research) has recently developed methods to match supply and demands for skills on an individual level (the "House of Skills: Perfect Fit" project), using big data to gain insights into possibilities for intersectoral mobility. Both projects stress the necessity of a universal skills language (taxonomy/ontology).

The *qualification structure* that SBB develops and maintains can play an important role in these projects.

## Projects and pilots

### Job Perspectives

One of SBB's central projects is Job Perspectives (Kans op Werk). By combining available sources, data collected by ourselves and insights from the centres of expertise in our network, we compute for newly starting students their prospects of finding a job fitting their education directly after graduation.

This project uses data from economic development prognoses by the Bureau for Economic Policy Analysis, from the Central Bureau of Statistics and the Employee Insurance Agency, and from vacancies advertised online. The Internet spider (Jobfeed) collects vacancies, which are then cleaned and coded (with jobs related to qualifications), and enriched with corrections gathered through a large-scale Internet survey among our 250,000 accredited work placement companies. This body of data is then combined with the predictions about how many students will pass their exams and enter the labour market.

The underlying data and outcomes are published for students,[38] and for the various other stakeholders, in a variety of reports and dashboards.[39]

The focus of the Job Perspectives project is mainly quantitative, and reports at the level of qualifications and jobs.

---

38  www.kiesmbo.nl [13 Feb. 2020].

39  e.g. https://www.argumentenfabriek.nl/ [9 Mar. 2020].

▶ **Figure 2.6. Sample screenshot from KiesMBO (TVET portal for study and career choice)**
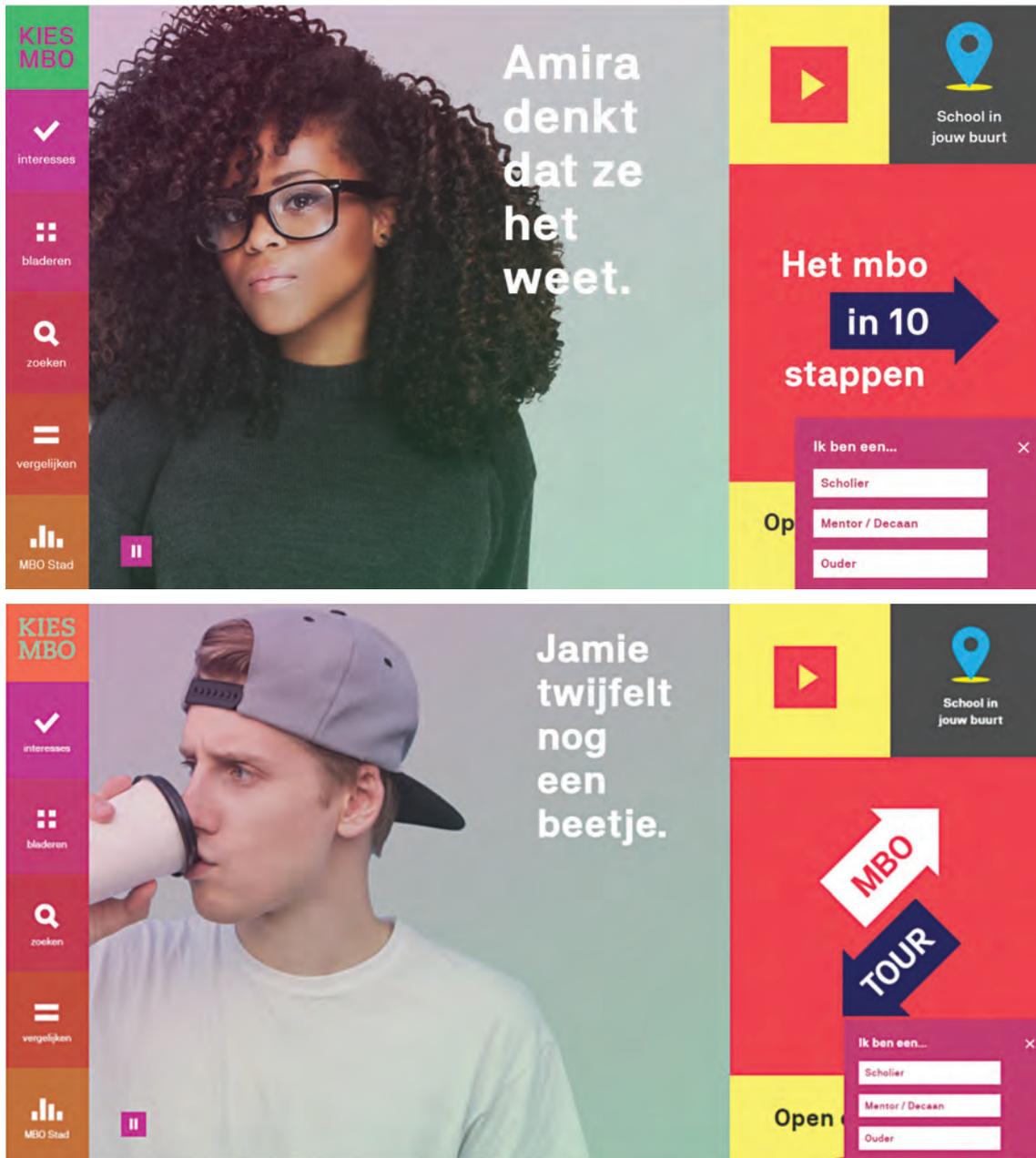


**Source:** https://www.kiesmbo.nl/. (28 Aug. 2020).

## Pilot projects

SBB has recently begun some pilot studies with a view to gaining information about changes and developments in job tasks at a more detailed level (knowledge, competencies, skills).

One such pilot study attempted to chart changes in the job content for two professions (process operators and intermediaries) by means of text analysis. The results are promising, though the human input required is still substantial.

In another study, we tested a tool developed by the Australian company Faethm, which shows and specifies the impact of technology on companies over periods between one and 15 years. The algorithms use the actual job content and the trends in 17 different technologies to compute the impact of technology on the workforce. The findings show how large the productivity gain from technology can be, and how many people are affected through possible reduction or addition of jobs. For the pilot study we applied the tool to a cohort of graduates entering the labour market in 2018. Thus we were able to make predictions for the number of people from that cohort who may become redundant and/or will have to reskill. The tool shows possible job corridors; it also shows which competencies, skills, etc. need to be improved to enable workers to transfer to other jobs.

Our preliminary conclusion is that, while these methods may be very valuable, the choice of taxonomies to be used is crucial. Together with our stakeholders, we are currently examining ways to transfer these kinds of results into action.

## 2.2.3. Conclusions

All in all, the potential of big data is promising. The greatest challenge lies in how to assess and use unstructured text data, such as vacancy advertisements or contributions on Internet platforms. Furthermore, work on big data (as well as "normal data") needs to be supported by and combined with accurate theory, and involve the full network of stakeholders.

## ▶ 2.3. Lessons learned from selected studies on education–labour market matching

**Gábor Kismihók**

### 2.3.1. Text mining in organizational research

Despite the ubiquity of textual data, to date few researchers have applied text mining in answering questions relating to organizational research. Text mining, which essentially entails a quantitative approach to the analysis of (usually) large volumes of textual data, helps to accelerate the acquisition of knowledge by radically increasing the amount of data that can be analysed. This article (Kobayashi et al., 2017a) aims to acquaint organizational researchers with the fundamental logic underpinning text mining, the analytical stages involved, and contemporary techniques that may be used to achieve different types of objectives. The specific analytical techniques reviewed are (a) dimensionality reduction, (b) distance and similarity computing, (c) clustering, (d) topic modelling and (e) classification. The article describes how text mining may expand the scope of contemporary organizational research by allowing existing or new research questions to be addressed using data that are likely to be rich, contextualized and ecologically valid.[40] After an exploration of how evidence for the validity of text mining output may be generated, the authors conclude the article by illustrating the text-mining process in a job analysis setting using a data set composed of job vacancies.

### 2.3.2. Text classification for organizational researchers: A tutorial

Organizations are increasingly interested in classifying texts or parts thereof into categories, as this enables more effective use of the information. Manual procedures for text classification work well for up to a few hundred documents. However, when the number of documents is larger, manual procedures become laborious, time-consuming and potentially unreliable. Techniques from text mining facilitate the automatic assignment of text strings to categories, making classification expedient, fast and reliable, which creates potential for its application in organizational research. The purpose of this research (Kobayashi et al., 2017b) was to familiarize organizational researchers with text-mining techniques from machine learning and statistics. The authors describe the text classification process in several roughly sequential steps, namely training in data preparation, pre-processing, transformation, application of classification techniques and validation, and provide concrete recommendations at each step. To help researchers develop their own text classifiers, the R code associated with each step is presented in a tutorial which draws from the authors' own work on job vacancy mining. The study ends by discussing how researchers can validate a text classification model and the associated output.

### 2.3.3. Automatic extraction of nursing tasks from OJVs

This study (Kobayashi et al., 2016) explores the use of text-mining procedures (Aggarwal and Zhai, 2012) to automatically extract a specific type of job information, namely tasks/activities from online nursing job vacancies. The method was developed following the approach outlined in Solka, 2008, consisting of text pre-processing, feature extraction, application of a classification algorithm and evaluation of classification. The OJVs were provided by an online recruitment agency called Monster.

---

40  Ecological validity refers to the ability to generalize study findings to real-world settings. High ecological validity means research findings can be generalized to real-world settings. Low ecological validity means they cannot.

### 2.3.4. Big (data) insights into what employees do: A comparison between task inventory and text-mining job analysis methods

While much work has changed over recent decades, detailed studies on what it is that employees do in these changing jobs are lagging behind. Researchers and practitioners alike lack information about how work could best be organized to facilitate employee well-being. The development of methods such as text mining could provide a viable means of addressing the difficulties associated with current methods of job analysis in attempting to uncover what employees do in the broader context of the changing nature of work. The findings of this study (Berkers et al., 2019) show that it is possible to automatically extract from OJVs and analyse tasks that have relatively high correspondence with tasks collected using a task inventory. The text-mining method generally performed well on average importance and inclusion ratings owing to the lower level of detail of the automatically extracted tasks. In addition, the text-mining method provided a wider variety of contextualized tasks drawn from a larger sample of jobs, whereas the task inventory yielded more detailed and more mundane tasks. Text mining thus complements rather than replaces current job analysis methods. In addition, the study showed that not all tasks were equally related to employees' job satisfaction, work overload and emotional exhaustion, suggesting that it is worthwhile to include more task data in studies on employee well-being.

### 2.3.5. Survey vs scraped data: Comparing time-series properties of web and survey vacancy data

This paper (de Pedraza et al., 2019) studies the relationship between a vacancy population obtained from web crawling and vacancies in the economy inferred by a national statistical office (NSO) using a traditional method. The authors compare the time-series properties of samples obtained between 2007 and 2014 by Statistics Netherlands and by a web-scraping company. They find that the web and NSO vacancy data present similar time-series properties, suggesting that both time series are generated by the same underlying phenomenon: the real number of new vacancies in the economy. The authors conclude that, in their case study, web-sourced data are able to capture aggregate economic activity in the labour market.

### 2.3.6. Combining learning analytics with job market intelligence to support learning at the workplace

Numerous research articles are concerned with the issues surrounding the deployment of e-portfolios. Without proper mentorship, well-designed e-portfolios and stable systems, the learner's experience is often negative. The authors of this chapter (Berg, Branka and Kismihók, 2018) review how two large-scale big data infrastructures – Jisc UK's national experimental learning analytics (LA) and Cedefop's European Job Market Intelligence (JMI) infrastructure – can be combined to provide optimized and just-in-time advice. LA is a new data-driven field and is rich in methods and analytical approaches. It focuses on optimizing the learning environment by capturing and analysing the learner's online digital traces. JMI digests vacancy data, providing a broad overview of the job market including new and emerging skill demands. The authors envisage a future in which e-portfolios are populated with authentic job market-related tasks providing transferable long-term markers of attainment. The e-portfolios are populated through entity extraction, performed by running ensembles of machine-learning algorithms across millions of job descriptions. The process is enhanced with LA, allowing us to approximate the skill level of the learner and select the tasks within the e-portfolio most appropriate for that learner relative to their local and temporal workplace demands.

# ▶ 2.4. Bridging the gap between skills and occupations: Identifying the skills associated with Canada's National Occupational Classification[41]

**Labour Market Information Council**

## 2.4.1. Overview, rationale and objective

▶ The skills required to succeed in today's world of work are rapidly changing. Workers experience pressure to continuously improve their skills, and employers struggle to find workers with the right skills to achieve their goals.

▶ Developing a pan-Canadian mapping system that links skills to occupations is an important step towards improving our understanding of the changing nature of jobs.

▶ A five-phase plan is proposed to assess, develop and maintain a mapping between the recently developed skills and competencies taxonomy of Employment and Social Development Canada[42] and the National Occupational Classification (NOC) system.

▶ The various approaches to achieving such a mapping will be evaluated against a number of established criteria, including, among others, data collection requirements, statistical rigour, utility in supporting people to make informed decisions, and the cost of applying and maintaining these different approaches.

▶ To ensure the skills and competencies taxonomy and its mapping to the NOC system continue to evolve to meet the needs of stakeholders, external input and feedback will be sought throughout the process.

▶ To ensure the credibility, rigour and integrity of the final mapping, Statistics Canada and Employment and Social Development Canada will manage and oversee the statistical infrastructure required to maintain and update the mapping.

## 2.4.2. Introduction

Current and future skill shortages in Canadian labour markets have been a major concern of policy-makers in recent years. There is a critically important need to better understand the underlying skills and training needs of employers and workers, and to help education and training providers better prepare workers for navigating the changing world of work and support them in doing so. Labour market experts have long called for improved clarity in the definition and measurement of skills and their relation to the job market. Canada, however, lacks an open, credible data source that contains reliable information on the skills associated with jobs. To address this deficiency, some relatively straightforward steps can be taken to leverage existing infrastructure, notably the NOC, and recent initiatives have sought to develop a richer skills and competencies framework. As part of these efforts, Employment and Social Development Canada (ESDC), Statistics Canada (STC) and the Labour Market Information Council (LMIC) have formed a partnership to fill this key LMI gap. Together, and with continuous stakeholder engagement, the three entities will work to create a shared and open Canadian framework linking ESDC's skills and competencies taxonomy (see appendix below) to the NOC system of occupations in a manner that conveys the evolving skill requirements of work. A statistically robust linkage – or mapping – between the well-defined ESDC skills and competencies taxonomy and the NOC system will enable us to better articulate the composition and distribution of skills across jobs (e.g. occupations and industries) and worker characteristics (e.g. their education level).

---

41 This section has been prepared jointly by the staff of the Labour Market Information Council, Statistics Canada (Labour Statistics Division) and Employment and Social Development Canada (Labour Market Information Directorate).

42 https://noc.esdc.gc.ca/SkillsTaxonomy/SkillsTaxonomyWelcome/f5ec9457d5a540eeb4529c2698acb19a [14 Aug. 2020].

> ▶ **Box 2.1. Skills measurement: Caveats and considerations**
>
> Any project to document the skills associated with the NOC will need to consider skills measurements and any trade-offs inherent in the various approaches. This is true for the measurement of both the supply of and the demand for skills. For instance, psychometric and competency-based skill assessments can lead to very granular insights at the individual level (the supply of skills), but these tests are costly to undertake, and there can be ambiguity in respect of what is being measured. Similarly, having job analysts monitor and evaluate the tasks of people in the 500 NOC occupations may yield accurate information in terms of skill requirements; but the process would be time-consuming, and any results would be at risk of being out of date by the time the information were validated and published. Other more aggregated or big-data-driven means to capture skills demand may be efficient and cost effective but lack specificity. These measurement considerations will be borne in mind, along with other criteria (discussed below), when evaluating the various approaches linking skills to the NOC.

**Source:** Author.

## 2.4.3. Background

In designing education, training and employment-support programmes, various levels of government and private-sector entities have developed a range of standards and frameworks to classify both workers and job characteristics, including skills, knowledge domains, tasks, etc. These classifications tend to be specific to the programme or policy being monitored or evaluated, or are the intellectual property of the private-sector firm in question.

With a view to the future, given the increasing importance of and focus on skills, we need better data and insights to inform our decision-making. The NOC system is the framework for describing the world of work in Canada and is the basis on which we should start to develop and add new LMI. In particular, we need to build on the established statistical system of occupations to structure and organize skills information. Such a system should also consider how to capture, directly or indirectly, the measurement of skills (see box 2.1). Leveraging the structure of the NOC system and linking NOC to the ESDC skills and competencies taxonomy requires a mapping from NOC-based occupations to specific skills. Such a mapping would connect each occupational category to a set of skills; in this sense, the mapping to skills would deepen the current NOC framework without changing its foundation. A robust and sustainable mapping of this nature will be a first step in helping to guide individuals, employers, education and training providers, policy-makers, researchers, career practitioners and others to a better understanding of the skills needed today and tomorrow.

## 2.4.4. A Canadian skills and competencies taxonomy

In an effort to link skills with occupations, ESDC has identified nine essential skills "for learning, work and life" and developed 372 essential skills profiles representing 361 occupations to inform training providers about skills needs and to better monitor skills development funding.  These skills, and the profiles, are being reviewed to determine how they can be modernized to better reflect changes in the labour market. There is also the ESDC career handbook, a more detailed source of NOC-based skills information (including 930 occupational profiles) that acts as a career-counselling component of the NOC. To enrich and complement these efforts, in 2017 ESDC began developing a skills and competencies taxonomy (see appendix below) that streamlines terminology across a number of domains and concepts. The taxonomy provides, in addition to a skills classification, a comprehensive framework to characterize both workers and jobs across seven other mutually exclusive categories (see figure 2.7). It is important that categories are exclusive to avoid duplication of terms and definitions. The skills component of the taxonomy includes 47 distinct skills (known as descriptors), each accompanied by definitions and organized into five skill groups: foundational, analytical, technical, resource management and interpersonal. The skills and competencies taxonomy was developed using a variety of internal resources, quantitative and qualitative research, and stakeholder consultations. An initial goal was to establish consistency in the way occupational and skills information was presented,

▲ **Figure 2.7. ESDC skills and competencies taxonomy framework**

| Skills | Abilities | Personal attributes | Knowledge | Interests | Work Context | Work Activities | Tools and technology |
|--------|-----------|---------------------|-----------|-----------|--------------|-----------------|----------------------|
| Foundational | Cognitive | *e.g. adaptability/ flexibility, attention to detail, cooperation, creativity / originality / innovation, dependability, independence, initiative, judgement, leadership, self-awareness, etc.* | Manufacturing and production | **Holland/ RIASEC** | Work Values | Information input | *Descriptors and descriptor subcategories to be defined* |
| Analytical | Physical | | Communications and transport | ▲ Realistic ▲ Investigative ▲ Artistic ▲ Social ▲ Enterprising ▲ Conventional | Physical demands | Interacting with others | |
| Technical | Psychomotor | | Law and public safety | **Canadian Work Preference Inventory (CWPI)** | Environmental conditions | Mental processes | |
| Resource management | Sensory | | Scientific knowledge | ▲ Directive ▲ Innovative ▲ Methodical ▲ Objective ▲ Social | Structural job characteristics | Work output | |
| Interpersonal | | | Social sciences and arts | | Interpersonal relations | Service/care provision | |
| | | | Education and training | | | | |
| | | | Health services | | | | |
| | | | Mathematics and science | | | | |
| | | | Engineering technology | | | | |
| | | | Business, finance and management | | | | |
| | | | Etc. | | | | |

**Note:** "Tools and technology was originally integrated under "Work context". In the current version of the taxonomy, this is a stand-alone category.

**Note:** The terms "skills", "ability" and "competency" are often used interchangeably in common language. However, the literature suggests nuanced differences. Competencies involve the use of skills, abilities and attributes to complete a task or successfully meet demands.

**Source:** Author.

while drawing on several rich sources of data to inform ESDC tools, including the essential skills profiles, the Career handbook and the skills and knowledge checklist available in Job Bank.[43] With respect to the skills category, the taxonomy leverages existing efforts in defining skills. The taxonomy builds on the O\*NET system, developed by the US Bureau of Labor Statistics.[44] It is also informed by pan-Canadian information sources, including Red Seal Occupational Standards and the national occupational standards. Following extensive reviews of these and other international frameworks, descriptors were identified, compiled, and organized into the continually refreshed and updated skills and competencies taxonomy. These well-defined categories and subcategories will help improve the definition and understanding of the skill in question.

## 2.4.5. Connecting the skills and competencies taxonomy to the NOC

Measuring and defining skills in the context of the world of work, from both the supply and the demand perspectives, is a complex process. One means by which we can improve LMI and insights in this respect is to build a mapping from the ESDC skills and competencies taxonomy to the NOC system. Such a mapping would essentially document the skills associated with each occupation. As noted above, ESDC, STC and LMIC are currently collaborating – along with their provincial and territorial counterparts – to implement a phased approach to the development and evaluation of such a skills-to-NOC mapping. A phased approach will ensure that the development, roll-out and maintenance of the skills mapping to occupations is statistically robust, methodical, client-oriented and operationally sustainable. The five phases of the mapping project are as follows:

1. ongoing consultation and improvement of the skills and competencies taxonomy;

2. identifying and evaluating mapping approaches;

3. pilot testing;

4. assessing and validating the pilot tests using pre-defined criteria;

5. dissemination, administration and implementation.

Although assessment and validation are the specific purposes of phase 4, each phase will be viewed through a critical lens. Throughout the process, stakeholders will be engaged in each phase of the project through direct consultations and requests for feedback. Each of the five phases is discussed in more detail below.

### Phase 1: Ongoing consultation and improvement of the skills and competencies taxonomy

The skills and competencies taxonomy in its current form is only the first iteration, presented for input and feedback. The taxonomy will remain evergreen as ESDC continues consultation efforts with provincial and territorial governments, national sectoral organizations, workers' organizations, educators and training providers, career practitioners and the private sector, as well as other federal agencies and external experts (see box 2.2 for more information on efforts related to digital skills). This process is important, not only to gain direct feedback on the skills and competencies taxonomy but also to enable the collection of more accurate and relevant descriptors.

---

43  See https://www.jobbank.gc.ca/home [27 July 2020].

44  O\*NET is an open data source that includes a skills taxonomy, variables describing work and worker characteristics, and a mapping to the Standard Occupational Classification (SOC).

▶ **Box 2.2. Digital skills**

Innovation, Science and Economic Development Canada (ISED) has created an interdepartmental working group on skills and talent to bring together representatives from across government to discuss a range of skills-related policy issues. A sub-working group has been established to study and advance concepts related to digital skills and help support policy, programming and statistical objectives in this field across government. The existing structure of the skills and competencies taxonomy defines digital skills as follows: "Understanding and using digital systems, tools and applications, and to process digital information". This definition is currently organized within the skills category and the foundational skills subcategory. Discussion is currently under way within the inter-departmental sub-working group on how to build on the digital skills framework as it pertains to the ESDC skills and competencies taxonomy. Specifically, the goal for 2019 was to focus on the study of digital skills, on the basis of nationally and internationally recognized standards, practices and concepts, and to reach consensus on a definition that meets the needs of the membership.

**Source:** Author.

## Phase 2: Identifying and evaluating mapping approaches

In phase 2, the partnership will research, identify and assess various methods for linking skills to all occupations in a consistent manner. The initial focus will be on the potential nature of the relationship between skills and occupations and the appropriate data collection methods required. There are a number of prevailing approaches, including, but not limited to, the following:

▶ consulting occupational experts;

▶ seeking direct input from workers and employers;

▶ obtaining input indirectly through online job posting/CV data; and

▶ hybrid approaches.

Within these broad approaches, methods currently used range from reliance on occupational analysts, as in the O*NET system, to quantitative methods preferred by private organizations. LinkedIn, for example, uses text supplied in job postings to identify in-demand skills, while Nesta, a UK-based innovation foundation, has built a list of skills in demand and linked them to occupations using machine-learning algorithms. These and other approaches to skills data collection and mapping to occupations will be explored during this phase of the project. It will be important at this stage to establish the criteria for evaluating the benefits and challenges associated with, and trade-offs inherent in, the various approaches. Table 2.3 summarizes seven broad criteria that will help in evaluating the approaches. For example, the use of occupational analysts to map skills to NOC is likely to be statistically sound and granular, but such approaches are costly to sustain and less responsive to the ever-changing world of work. Conversely, using data from online sources is inexpensive to sustain, is highly flexible and provides insights in (almost) real time; however, the statistical robustness of using only online data is questionable. It may be possible to identify a hybrid solution that draws on the best of a variety of methods. One example of a hybrid method is a baseline skills-to-NOC mapping based on expert analysis, regularly adjusted in a robust manner to real-world signals that might be drawn from online data.

There are also considerations regarding how the link between skills and occupations is defined, e.g. the importance of a skill for a job, or the complexity of using that skill in the job. For example, in the O*NET system, skills are mapped to occupations through an importance rating on a scale from 1 to 5, where 1 represents "not important" and 5 "extremely important". ESCO, on the other hand, maps skills to occupations through binary classifications – a skill is either "essential" or "non-essential" for the job. Evaluating the approaches to mapping skills to NOC will also involve careful consideration of the manner in which the relationship is being mapped.

▶ **Table 2.3. Key criteria for evaluating the mapping project**

| Benchmark | Description |
| --- | --- |
| Flexible | Managed and executed in a way that enables it to be modified, augmented or adapted to respond to changing labour market conditions and to capture emerging skills |
| Sustainable and cost effective | Adequate resources to maintain and update the mapping |
| Representative | Reflects the different ways employers, workers and training providers express skill requirements |
| Granular | Incorporates greater specificity of skills and occupation-specific data |
| Responsive | Enables policy-makers, career and employment counsellors, curriculum developers and others to make better-informed decisions about skills training and education |
| Measurable | Allows for the reasonable measurement of skills |
| Statistically sound | Sound empirical techniques ensure the resulting estimates of skills levels and distributions are representative of Canadian labour markets |

**Source:** Author.

## Phase 3: Pilot testing

Following an evaluation of these various approaches and selection of the recommended potential way forward, a set of occupations to use in a pilot study will be identified from the NOC. The number of occupations used in this pilot testing phase will depend on which methodology has been selected in phase 2. More qualitative and labour-intensive methods would suggest that a small subset of the NOC list be used, whereas highly automated methods could be easily scaled to a large number of occupations.

## Phase 4: Assessing and validating the testing results using pre-defined criteria

After pilot testing, further information and data will be gathered from appropriate sources to address issues that are likely to arise during the testing phase. It will be important at this point to conduct a quality control review to ensure compliance with a set of predetermined validation criteria. At this stage, and indeed throughout the life cycle of this project, we will assess whether the skills-to-NOC mapping is on track to meet the key criteria set out in table 2.3 above.

## Phase 5: Dissemination, administration and implementation

The final phase of this project is the publication and promotion of the skills and competencies taxonomy and its mapping to occupations. At this phase, key considerations are credibility, accessibility and capacity to evolve. First, with respect to credibility, the mapping and the taxonomy will be managed and overseen by STC and ESDC. Second, as with the NOC system, the skills and competencies taxonomy and its mapping to the NOC will be a public good that will be open and available online in a machine-readable format, accompanied with clear and complete descriptions of the methodology and metadata. Third, both the taxonomy and its mapping to the NOC must be flexible and responsive to the needs of stakeholders and the availability of new technologies.  Evidently, changes to the taxonomy will require revisions to its mapping; however, we should not assume that the approach to mapping initially selected will necessarily remain the best option available. As new technologies and data emerge, so the infrastructure established to maintain an evolving

skills and competencies taxonomy and mapping to occupations must respond with new analyses, pilot testing and, ultimately, implementation of more efficient and robust techniques.

Finally, the scope and scale of this endeavour should not be underestimated. It will be crucial to establish a clear operational plan – along with costs and responsibilities – that ensures this project is sustainable in the long term. The skills profiles of Canadian workers and the skills requirements of Canadian employers will continue to be essential pieces of LMI well into the future. The LMIC is committed to ensuring that Canadians will have access to the skills information and insights they need and want.

## 2.4.6. The way forward

Jobs are evolving rapidly as workplaces innovate and adopt new technologies. The skills required for these jobs are shifting at a similar pace. In an effort to improve our understanding of the skills associated with jobs, a phased project is under way to evaluate and develop a pan-Canadian mapping to link skills to occupations. This is an important first step towards improving our collective understanding of the relationships between skills and jobs; but it is only one step, complementary to other initiatives whose aim is to ensure that, in a changing and dynamic world of work, Canadians have the right skills to succeed and employers have access to the right talent to grow their businesses. The aim of the LMIC is to ensure that the skills and competencies taxonomy and its association with the NOC are transparent and accessible to all. This is particularly relevant given the varying, often conflicting or inaccurate, definitions of skills, and multiple mappings that charge fees for access. In this regard, a key objective is to build alignment towards a widely recognized taxonomy – or at the very least to ensure some general convergence – when speaking about skills and occupations. Finally, the LMIC will demonstrate accountability and transparency by engaging with its partners and stakeholders, disclosing its progress throughout the project life cycle, and providing information that is timely, accurate and relevant to the wider public. The success of this project rests fundamentally on continual engagement with all actors through an open and inclusive consultation process.

## Appendix: ESDC's skills and competencies taxonomy

In order to complement other federal, provincial and territorial employment programming efforts around skills identification and utilization, ESDC has developed a skills and competencies taxonomy to help facilitate a pan-Canadian dialogue on skills. The taxonomy serves to streamline terminology across a number of competency domains and concepts (e.g. skills, personal abilities and attributes, knowledge, interests), occupational work context, work activities, and tools and technology information, while aiming to improve the comparability of their incidence and application throughout occupations and sectors. The taxonomy also complements ESDC's development of a range of LMI products, such as the Canadian skills profiles, which will detail the competency requirements for entry into specific occupations, as well as provide other indicators relating to the use of skills (e.g. the importance of a skill within a particular occupation, and/or the frequency of its use). The taxonomy was constructed on the basis of internal products (e.g. the Career handbook, skills and knowledge checklist, and essential skills profiles) as well as a variety of national and international competency-based frameworks, including the US O*NET system (see table 2.4). ESDC continues to consult with internal and external stakeholders, including the provinces and territories, in order to validate and improve the content of the skills and competencies taxonomy.

### Note on the nuance between "skills" and "competencies":

The literature on skills and competencies suggests a nuanced relationship between the two concepts, namely, that competencies involve the use of skills, abilities and attributes to complete a task or successfully meet demands. Consistent with the existing literature, ESDC proposes the following definitions for the skills and competencies taxonomy:

▶ *competencies:* the combined utilization of personal abilities and attributes, skills and knowledge to effectively perform a job, role, function, task or duty;[45]

▶ *skills:* developed capacities that an individual must have to be effective in a job, role, function, task, or duty;[46]

▶ *personal abilities and attributes:* inherent and developed aptitudes that facilitate the acquisition of knowledge and skills to perform at work.[47]

## Note on digital skills

One of the outstanding challenges of the taxonomy is the integration of the digital skills concept. ESCD is undertaking a number of activities to ensure this concept is properly integrated in the taxonomy, for instance performing the secretariat role for the sub-working group on classification of skills and competencies of the ISED–ESDC–Statistics Canada working group on skills and talent. The sub-working group is currently exploring the terms, definitions and placement of digital skills in the taxonomy (see also box 2.2).

▶ **Table 2.4: Sources for ESDC's skills and competencies taxonomy**

| Framework / Report / Article title | Author(s)/disseminator | Code |
|---|---|---|
| Global Competency Framework | Council of Ministers of Education, Canada (CMEC) | **CMEC** |
| The definition and selection of key competencies: Executive summary | Program for International Student Assessment (PISA) | **PISA** |
| Program for the International Assessment of Adult Competencies (PIAAC) | PIAAC | **PIAAC** |
| International Symposium 2017 – Canada Paper | International Centre for Career Development and Public Policy (ICCDPP) | **ICCDPP** |
| 21st Century Competencies | Ontario Public Service – Learning Partnership | **ON** |
| Atlantic Canada Framework for Essential Graduation | Atlantic Provinces Education Foundation (now Council of Atlantic Ministers of Education and Training – CAMET) | **CAMET** |
| "Potential hires coming up short in 'soft skills', employers say" | Canadian Broadcasting Corporation (CBC) | **CBC** |
| O*Net | National Center for O*NET Development (US Department of Labor) | **O*NET** |
| Ontario Skills Passport (OSP) | Ontario Ministry of Education | **ON EDU** |
| New Work Smarts and New Work Order Reports | Foundation for Young Australians | **FYA** |
| COPS/CAPS/COPES-Interest Inventory, CAPS-Abilities, and COPES-Work Values[a] | COPS/CAPS/COPESystem (EdITS) | **COPS/ CAPS / COPES** |
| British Columbia (BC) Public Service Competencies | BC Public Service | **BC** |
| *Career handbook* (CH), skills & knowledge checklist (S&K), essential skills (ES) | Employment and Social Development Canada | **ESDC: CH, S&K, ES** |
| Employability Skills 2000+ and General Innovation Skills Aptitude Test 2.0 | Conference Board of Canada | **CBoC: ES2K/GISAT** |
| Hazards Database | Canadian Centre for Occupational Health and Safety | **GoC: CCOHS** |
| Nos Compétences Fortes | Institut de coopération pour l'éducation des adultes | **ICÉA** |

---

45 Adapted from the International Society for Performance Improvement and the Organisation for Economic Co-operation and Development

46 Adapted from the US O*Net definition of skills.

47 Adapted from the US O*Net definition of abilities and work styles.

| Framework / Report / Article title | Author(s)/disseminator | Code |
|---|---|---|
| Classification of Instructional Programs | Statistics Canada | **Stats Can** |
| "It's human skills – not technical skills" | Globe and Mail | **GM** |
| "Let's stop calling them 'soft skills'" | It's Your Turn | **YT** |
| Occupational Information System Project | US Social Security Administration | **SSA** |
| United Nations Standard Products and Services Code | United Nations Standard Products and Services Code® (UNSPSC®), managed by GS1 US™ for the UN Development Programme (UNDP) | **UNSPSC** |
| "Making vocational choices: a theory of careers" (complete version not available online) | John L. Holland | **Holland** |

**Note:** [a] COPS = Career Occupational Preference System (an interest inventory assessment for career exploration and counselling); CAPS = Career Ability Placement Survey; COPES = Career Orientation Placement and Evaluation Survey.

**Source:** Author.

# ▶ 2.5. Bringing traditional sense to the big data craze: Emsi UK[48]

**Andy Durman**

## 2.5.1. Introducing Emsi

Emsi's mission is to drive economic prosperity through the effective use of LMI to better connect three audiences: people who are looking for good work; *employers* who are looking for good people; and *educators* who are looking to build good course programmes and engage students.

It is of critical importance that these connections are forged and developed within local economic ecosystems and regional economies. Accordingly, through our unique integration of a variety of labour market data sources, we create a detailed and holistic understanding of workforce supply and demand down to the local level. We use this intelligence, combined with our empirical labour market analysis and data expertise, to support a range of organizations as they seek to make better, more informed decisions relating to their core labour market.

To deliver on this mission, Emsi primarily serves three core types of organization, each with a common need for good, localized labour market insight, but each with different uses and applications for this intelligence:

▶ **education organizations** (e.g. colleges and universities), which need to align their curricula with current and emerging labour market and employer demand, while also informing their engagement with both students and (current and potential) employers through a good understanding of the connection between the education offer and labour market opportunities;

▶ **economic development organizations** (e.g. city regions), which need to shape local interventions to nurture economic growth – such as skills planning, business attraction strategies, etc. – by playing to regional skills and workforce strengths, and mitigating economic weaknesses and challenges;

▶ **employment-focused organizations** (e.g. staffing agencies and large corporate entities), which need to identify and quantify local talent pools and build knowledge of key labour dynamics that affect hiring and access to talent in order to mould workforce planning activities appropriately.

Emsi's value lies in its data, which we serve to key decision-makers and analysts within these organizations through web-based subscription software services, API data integration services or project-based consulting services.

Emsi is headquartered in the US state of Idaho, but today also serves its 2,000-strong global customer community from bases in Texas and in the British town of Basingstoke. In 2018 Emsi was acquired by the Strada Education Network, a US-based non-profit-making organization with a mission to improve lives by forging clearer and more purposeful pathways between education and employment.

## 2.5.2. The "why" and "how" of Emsi LMI

### The global talent challenge

The global economy and labour market have entered a period of profound and rapid change, fuelled by technological advance. This environment, often referred to as the fourth industrial revolution (or Industry 4.0), is driving significant shifts in the globalization of work and employment.

---

48 The following is a summary of key points presented to the ILO big data workshop by Andy Durman, Managing Director, Emsi UK. To find out more about Emsi's work, please visit www.economicmodeling.com (US) or www.economicmodelling.co.uk (UK) [13 Feb. 2020].

Workforce dislocation, through outsourcing to regions with lower labour costs, and workforce relocation, through moving operations to regions with specific expertise networks and lower costs, are creating job specialization and redistribution at a global level. These processes in turn are generating an ever-evolving need for in-country re-employment and training to ensure that populations can respond to changes in labour market demand, with the necessary skills to embrace new and emerging employment opportunities and to adapt to reductions in demand for the jobs they had previously done.

This is not the place to attempt an in-depth assessment of the specific disruptions that are occurring across the world's economies (for a more detailed exploration, see WEF, 2018b); rather, the intention here is to set out in brief the context for the need for good, detailed intelligence and data about the labour market at levels of granularity that really can convey the nuance of the changes that are occurring. Such data is critical to gaining, as rapidly as possible, an understanding of current and likely future trends, of the factors at play, and of the resulting changes in skills and employment demand, with the ultimate goal of helping organizations and individuals to make well-informed decisions to maximize the opportunities and mitigate the risks such disruption will present.

## Starting with structured data

Emsi's goal is to build a data set that illuminates as many facets of labour trends and skills demand as possible, providing not just a holistic overview of historic employment trends, but also contemporary and emergent details about employment at the skills level.

Emsi starts by building data from official, structured, labour market data sources. These are sources that are systematically collected and collated for the purpose of understanding the labour market, and are packaged in a neat and tidy structure. Such data is often driven by large-scale governmental statistical surveys and official data returns. The great strength of structured labour market data sources is that they typically capture the breadth of employment in the labour market, certainly at the employee level. In addition, the methodologies employed to categorize and structure the data provide consistency and comparability across in-country geography and over time. The use of standardized classifications also supports interoperability and connectivity between different data sets, facilitating a holistic overview of the labour market.

There are, however, a number of limitations inherent in such data that need to be considered. Given the nature of how data is collected, tested and published, there is often a lag (sometimes of several years) between the time at which data is collected and the time at which it becomes available for use. Using structural data, therefore, often means observing historic trends. Owing to the size and cost of producing such information sources, data is not updated very regularly, sometimes only on an annual basis (or decennially in the case of most major population censuses). Moreover, the coding systems applied to the data are updated even less frequently (often contemporaneously with significant data update processes such as a major population census). While all employment is still captured, a number of categories deteriorate in value and relevance as the nature of labour markets change; these would benefit from more frequent review, often through subdivision into more meaningful and more detailed groupings of job roles.

Finally, there are inherent limitations in data coverage and quality that also need to be considered when using such structured sources. Owing to concerns about confidentiality, data details that may be traceable to individuals (for example, when questions return very small numbers, often at the local level) are either omitted or rounded. But perhaps more pertinently, there are questions about the quality and coverage of the data sources themselves. While most developed nations have robust and well-funded statistical services (such as the Office for National Statistics in the United Kingdom and the Bureau of Labour Statistics in the United States), many nations, notably those in developing regions, lack the quality of infrastructure needed to produce trustworthy and complete data from which reliably useful insights can be derived.

## Harvesting big data

In addition to harnessing structural labour market information, Emsi also builds labour market insights from a wealth of big data – data that is captured from transactional activities and collated on an ongoing

basis. This includes data harvested from the wealth of intelligence contained within online job postings, as well as insights gleaned from worker profiles and CVs obtained through opt-in sources.

The advantage of data obtained from such sources, notably that derived from job postings, is the level of detail and granularity it offers, well beyond that available from standard structural data classifications: this covers specifics such as job titles and (hard and soft) skills, and can even flag up hiring activity in specific companies and locations. Collation of data can also take place in something approaching real time, rather than being constrained by the infrequency and publication lag of structural data sources. In order to benefit from such data, of course, it is necessary to have taxonomies and data collation processes that can translate raw, textual data into appropriate categories; but, provided such systems are in place, such data does enable the exploration of emergent skills and labour trends. Data of this more granular kind also enables more flexible searching and grouping of specified labour market characteristics.

Moreover, while reliance on structural labour market data sources requires there to be a robust statistical structure in place at government level, the nature of big data insights gleaned from transactional data means that intelligence can be derived from regions where structural labour market data is limited, unreliable or simply non-existent.

However, as with structural labour market information, big data sources are also subject to a number of limitations and weaknesses that must be considered when assessing the insights generated. The very transactional nature of the job postings used to generate insights means that they typically capture areas of the labour market that are churning (where hiring is taking place) and do not necessarily portray the whole of the labour market. Nor do these sources offer knowledge of the trends associated with the incumbent workforce; to fully understand these, we require other sources (such as worker profiles). In addition, as this data is sourced from searching the Internet, the big data approach captures only those jobs and skills that are being advertised through OJVs, which is only ever going to be a subset of all hiring activity (with a bias towards certain types of labour recruitment over others), even if the technology used can find all postings – which can by no means be assumed. While a large proportion of jobs are advertised through established recruitment portal channels, more and more job advertisements are being channelled through social and professional networks, and may not take the traditional form of job postings that can be harvested.

Quantification of demand from job postings is made more complicated by the nature of how volumes of job postings relate to actual employment volume. It is commonplace for the same role to be advertised through multiple channels to maximize visibility, especially for those jobs that are harder to fill. Similarly, a single job posting could in fact be used to hire multiple individuals for identical roles.

The final key challenge to consider is one of data quality, both in terms of what information is written into the posting in the first place, and in terms of how that data is processed and structured to make it useful for analytical purposes. On the first point, as there is no set structure or content requirement for a job posting, the nature of content will vary greatly across a large sample. The posting is the opportunity to advertise what an employer is looking for in a limited amount of space, and will emphasize the most critical requirements of the role and/or the hardest-to-find skills in the marketplace. It will not be an exhaustive list of skills requirements; often, basic skills are taken for granted or implied in the nature of the role being advertised. There may or may not be supplementary information about the wage on offer or the name of the employer (especially if the posting originates with a recruitment company). Location of employment can also be open to error and bias, particularly in areas surrounding large economic centres: for example, many job postings in the region around London are often simply stated as located in London to attract the biggest pool of candidates.

Thus, while big data sources, such as online job postings, undoubtedly present an opportunity to illuminate labour market and skills trends to a greater level of granularity and insight than ever before, such data does need to be treated with great care. This is starkly illuminated in a case study from Canada in 2014 (Curry, 2014). Presented with evidence of a growing skills gap across the nation, the Canadian Government passed major reform measures, including opening the doors to higher levels of immigration – only to find that this resulted in an increase in unemployment. A jobs report from Finance Canada, using data derived from job postings, was showing a vacancy rate in the region of 4 per cent and rising, in stark contrast to Statistics

Canada's own vacancy rate of less than 2 per cent and falling. It emerged that one of the key sources for the job postings data was Kijiji, a classified ads website that contained a large volume of spurious and anonymous job postings (as well as many genuine postings). Once this source was removed, the vacancy rate derived from job postings dropped to something very close to the rate derived by Statistics Canada.

## Combining structural and big data sources

It is clear, then, that both structural and big data sources of labour market intelligence have considerable strengths, but also limitations that are important to consider when analysing insights and applying them to decision-making. A number of those key limitations can be mitigated to some extent through good modelling and data-processing techniques. For example, in relation to structural data limitations, Emsi has developed a robust model to manage data suppressions and fill gaps in the original data sources. Similarly, Emsi's projection methodology enables historic trends to be turned into trend-based projections which evolve as new and updated data inputs are published.

On the job postings/big data side, limitations such as posting duplication and data structure are mitigated by the development and ongoing improvement of core taxonomies and textual recognition processes that convert raw text into structured and categorized information. For example, we have processes in place that are designed to de-duplicate the data to ensure that job adverts are not counted multiple times.

However, many limitations on both data types are inherent in the nature of the sources, and difficult to remove in isolation. This is why Emsi connects structural and big data sources together into an integrated flow of intelligence. Structural LMI provides a top-down, holistic view of how the building blocks of a labour market can be quantified. This data is typically used for general planning purposes, and enables a focus on the areas of labour market change that are having greatest impact. The big data derived from job postings, on the other hand, provides a bottom-up view of aspects of labour demand, furnishing a level of detail that can shape more tactical action. This data, which ultimately is a sample of hiring activity, is best suited as indicative, and is more about adding context to wider trends.

As such, the structural data supports a focus on key growth industries or occupations within a region that may warrant further exploration. The job postings data then helps to dig into this focus area to get a feel for more specific insights around the requirements of employers that are hiring, or specific skills that contextualize the underlying trend.

## Building a skills-based science

The challenge posed by the rapid labour market evolution currently under way is how to go beyond traditional, structural classifications that are too broad and do not keep pace with labour market change, to embrace a more detailed skills-based approach. But diving in at the skills level is a complex matter. In particular, the volume of different skills that may be identified can in fact present too granular a view of things for practical usefulness. While identification of very specific skills is needed in order to create the building blocks on which to found a new understanding of the labour market, it is often the case that practical usefulness and decision-making capability come from understanding how skills commonly *cluster together*, rather than simply identifying individual skills in their own right.

To support this work Emsi has built Emsi Skills[49], an open library of skills derived from the millions of postings, CVs and profiles sourced and categorized into Emsi's big data holdings. As well as providing the underpinning taxonomy for all of Emsi's big data processing, this library is intended to aid other agents and actors in the labour market (such as educators, economic developers and employers) by offering them a common skills language from which to work and around which to connect thinking and activity. The library captures approximately 30,000 of the most common skills emerging from this analysis and is updated regularly to reflect the constant evolution of skills demand. In an extension of this data-derived approach, the open data community are shaping the direction of the library to keep pace with how they see and feel the labour market developing from their experience on the ground.

49  https://skills.emsidata.com [15 Feb. 2020].

Having established the granular building blocks of the skills library, Emsi's economists and data scientists have developed a skills clustering methodology that visualizes and connects skills that are commonly found to cluster together in the labour market. Skills are mapped on the basis of their **prominence** – the frequency with which skills appear across job postings and profiles; their **correlation** – the uniqueness of the skills across postings and profiles; and their **overlap** – the frequency with which the same skills appear across different types of postings and profiles (those most often appearing in this way are often considered transferable skills). This visualization identifies skills that are most closely connected to each other, as well as identifying areas in which skills groupings interconnect and overlap.

The combination of the skills library and the clustering methodology that sits across the top of it enables the interrogation of labour market trends on a highly flexible basis, taking account not just of a broad range of occupational areas, but also of the geographic variations within these.

For example, in a study of the manufacturing sector across the United States, different core manufacturing clusters came to prominence in different states. Michigan, with its famed auto-manufacturing sector, saw vehicle manufacturing as its most prominent cluster, illuminating such skills as machining, producing hydraulics and building engines as specific manufacturing skills. In Tennessee, meanwhile, which is an emerging manufacturing base, traditional manufacturing is the most prominent cluster, including a number of niche welding skills.

In another example, a study of demand for data-science-focused roles in New York City and Washington, DC, flagged up a core set of common in-demand skills, such as software and app development, and business intelligence skills, but also some very regionally specific strands of skills demand. While New York registered a bias towards financial services, Washington showed a bias towards social science research. And even within the common skills areas, such as software and web app development, there was variation in the specific skills demanded – New York highlighting JavaScript and Agile, while Washington highlighted UI/UX design.[50]

## Going global

The methodologies used to harness both structural and big data labour market sources, and critically to connect them together to build a more holistic and joined-up picture of demand and supply, have been honed from Emsi's work in the United States, United Kingdom and Canada. These markets offer a wealth of data sources on a very large scale, and robust structural data upon which to develop and test sound methods and processes. Having established these approaches, Emsi is now applying the same principles in extending its reach globally, including to regions that have a less robust data infrastructure on which to build.

The Emsi Global project has focused on building regionalized data sets (at the level of the city and its environments (agglomeration)[51]) that combine government data with big data derived from worker profiles to map a high-level view of labour supply.[52] In launching the project, Emsi has drawn on the Emsi Skills library to develop a consistent transnational view of labour in 23 key clusters for which demand is strongest in multinational workforce planning activities (notably in technology, engineering, sales and marketing, office professionals, and health and science), across 35 countries.

While aimed primarily at a workforce planning rather than a skills planning audience, Emsi Global is creating a platform upon which a skills-based view of global labour markets can be built. Plans for further development include integrating data for more countries (where viable data can be sourced), opening up further categories of labour, and integrating intelligence derived from job postings intelligence to better reflect demand and competition for labour.

---

50  For more examples of this methodology in action, see the following research reports: www.economicmodeling.com/manufacturing-is-not-dead/; www.economicmodeling.com/data-science-research/ [15 Feb. 2020].

51  This relates to the "Principle Urban Area" – a city and its environs. Administrative geography was used to define these areas (e.g. NUTS3 for European countries)

52  www.economicmodeling.com/global/ [15 Feb. 2020].

### 2.5.3. Some observations

Exploring the labour market at the micro level of skills is creating an unprecedented volume of intelligence and insight about the drivers of changing skills needs, at a level of granularity that reflects the actual functioning of the labour market. The notion of occupations has made way for that of skills and skills clusters, and big data methodologies have enabled analysis to catch up with the real world. This new level of labour market categorization, for example, is already feeding into university course design at the modular/skills level and helping not only to shape programmes to meet specific and local labour market demands, but to help build a language that connects students with real employment opportunities at a skill level.

This process is, however, presenting a new set of challenges, which may be summarized as follows:

▶ **Big data doesn't have all the answers.** Anchoring big data back in structural information can help to highlight inherent limitations in the data, create a focus on what is most important and build up the wider picture.

▶ **Decision-makers face greater complexity.** With a greater level of data complexity comes a challenge to support and empower the decision-makers who are using the data. Great attention needs to be paid to the visualization of such intelligence and its integration into core decision-making workflows to maximize its use in practice. Alongside this, commensurate effort needs to be put into education to illuminate the data limitations that need to be considered when interpreting findings.

▶ **Understanding demand is only half of the story.** Having created a new language of skills demand, a similar development is required in articulating skills achievement to ensure that the skills system matches how employers seek talent. Most major education systems tend to work at the large qualification level, which was best suited to connecting with the job market at the occupational level. What we are now beginning to see is the emergence of "micro-credentials" and "digital badges" to facilitate a clearer connection at the skills level. This development can support an understanding of skills development within larger qualifications, and also help in creating more specific building blocks for upskilling.

# 3

# Use of big data by emerging and developing economies

This chapter discusses the challenges in analysing skills demand in developing countries by using big data analytics. Experiments of using LinkedIn data in Latin America, the Caribbean and India identify the limits as well as the potential. Big data analysis of skills demand in Myanmar identifies opportunities to supplement poor regular statistics and add value to the available labour market information.
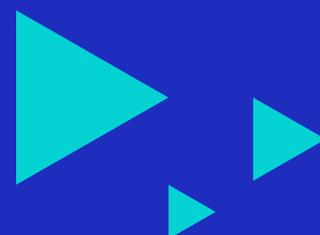
## ▶ 3.1. Viewing changes in skills demand in Latin America and the Caribbean through LinkedIn[53]

**Carlos Ospino Hernández**

### 3.1.1. New data sources to meet new challenges

Labour economists have traditionally used survey data from household surveys, enterprise surveys and administrative records. However, in today's environment they are under pressure to go out of their comfort zone and enter the world of large, unstructured data: big data. As a way into this arena, we focus on the professional networking app LinkedIn, used by millions of workers and jobseekers around the world. What makes LinkedIn's data quite unique is the fact that it captures information from the supply side of the labour market. LinkedIn's members post information about their current and past positions, their skills, their education and the work they do. We believe that the site is a new source of data that offers insights about the labour market that will be very useful for policy development.

What if you had the information to identify the fastest-growing job category in your country, or even in your home town, and the skills you need to learn to get hired for one of these jobs? And what if this data could also help you decide which jobs you could most easily transfer to with just the skills you have today, in case you wanted to change career or thought your current job might be automated soon? Obtaining the information to answer these questions is a primary concern for policy-makers, educators, students and workers everywhere at a time when the exponential growth of digital technologies, combined with the rapid development and deployment of robotics, AI, the Internet of Things and new platform technologies like LinkedIn are accelerating the pace of technological change and creating important shifts in the workforce.

Obtaining better and more timely insights into the changing demand for skills and occupations has become a pressing challenge for all countries. However, despite a global debate about the future of work and the changing demand for skills, few studies have been able to capture skills-level data in a way that is dynamic, cost effective, and reflective of the different markets, industries and occupations that characterize different geographical localities.

As we head into the fourth industrial revolution, high-skill employment may be increasingly affected. If history serves as an example, the destruction of jobs by technological advance will be accompanied by the creation of new ones in existing occupations and in other occupations that are still hard to imagine.

Technology enables the emergence of new sources of information, such as professional networks or job portals. These new sources can help us understand which skills different occupations require, how these requirements are changing, and how people can transfer from one occupation to another on the basis of these requirements – a sort of employment GPS.

### 3.1.2. Using LinkedIn data to investigate changes in skills demand

Using information available from LinkedIn profiles as a new source of dynamic labour market data on occupations and skills, we can provide new evidence to characterize changes in skills demand associated with shifts in occupations. A unique feature of LinkedIn's data is the availability of granular measures of skill importance by country and occupation. This data allows us to examine how similar occupations may differ in their skills composition across a range of countries, and to measure the corresponding shifts in skills demand associated with changes in occupations for each location. While the results are only representative

---

of the subset of workers with LinkedIn profiles, they provide an important perspective on an increasing, and arguably very relevant, portion of the labour market.

These data can also help in considering the degree of transferability of workers between declining and emerging occupations. Understanding the transferability of skills is one of the most promising areas of analysis that can be undertaken using LinkedIn data. Throughout history, the development and expansion of new tasks and occupations have helped to outweigh job losses caused by waves of automation. However, this expansion can be slowed – and transitions could be more costly – if recent graduates and/or workers who have lost their jobs as a result of automation do not have the skills required to perform these new tasks or occupations. Possessing and promoting more transferable skills – i.e. those skills that are important to and shared across different occupations – may help individuals to better withstand labour market disruptions in a dynamic digital economy.

We assess the transferability of workers employed in declining occupations to expanding portions of the economy as a first step in identifying the set of policies that may be needed to accelerate reallocation and economic adjustment, and to create more resilient learning and labour pathways for individuals. We leverage LinkedIn's granular skills data to create a distance measure that estimates how close two occupations are, based on the skills that workers in those two occupations share. Potential uses of this analysis include identifying those occupations into which workers in declining occupations can transfer, identifying alternative career paths for workers wishing to switch occupations, and identifying transitions with high potential for employment growth.

## 3.1.3. Using LinkedIn to explore trends in Latin America and the Caribbean

LinkedIn is one of several non-traditional data sources that can help explain trends in both occupations and relevant skills. LinkedIn provides accurate data on a subset of the labour market in Latin America and the Caribbean encompassing certain economic sectors and knowledge occupations (both highly automatable and not) that require higher education. This is because LinkedIn users in this region are, in general, more highly educated than average, with most having a partial or completed tertiary education.

Comparing the number of members in this professional network with the labour force aged between 15 and 64 in 18 countries in Latin America and the Caribbean for which we have data in the IADB's Labour Markets and Social Security Information System, we estimate that LinkedIn members represented 20 per cent of the region's labour force in 2017. In recent years, LinkedIn membership has significantly increased in the region, where the number of profiles has now risen to over 86 million. Comparing the data on the average age of the LinkedIn members with ILO data on employment by age and economic activity, a World Bank study (Zhu, Fritzler and Orlowski, 2018) found that with LinkedIn members are older on average by five years. Also, the proportion of women registered with LinkedIn is slightly greater than that in the ILO data.

Globally, the sectors most strongly represented in LinkedIn are information and communication technologies (48 per cent); occupational, scientific and technical activities (26 per cent); mining (25 per cent); insurance and finance activities (22 per cent); art, entertainment and recreation (14 per cent); and manufacturing industry (3 per cent). In Latin America and the Caribbean, the most represented industries are information and communications, followed by mining. Although information for other economic sectors exists, the sample size does not allow to include them in the comparison between LinkedIn data and IADB harmonised household data.

Our analysis yielded several key conclusions and recommendations:

▶ Across all the countries examined, technology-related occupations, such as software development, as well as advanced digital skills, like web development tools, are on the rise. Education and training providers must adjust their curricula and supply of courses to ensure that learners are acquiring a command of the specific tools that are most in demand and train professionals who can meet this growing demand. This may imply shifting some resources from administration/management courses towards courses that teach advanced digital skills.

▶ People-centric roles are also growing. Many of these occupations are highly skilled and have increasingly important digitally oriented components, but also require high levels of social intelligence to gauge and elicit people's reactions and to make high-level decisions on the basis of complex information. At the same time, lower-skill service-oriented and caretaker occupations are also emerging in several countries. All of these appear to be the least likely occupations to be automated.

▶ Countries with more connected networks of occupations may have a better chance of helping workers to make the transition out of declining occupations. Our results showed that in countries where occupations are closely related, workers in declining occupations have a wider array of options.

These results are also pertinent to analysis of the occupational transitions that yield the highest gains in hiring rates. Identifying the linkages between occupations requires access to granular data on skills and occupational characteristics, in order to construct measures of occupational relatedness. Fortunately, these data are becoming readily available to policy-makers.

However, the analysis presented here needs to be interpreted with care. On LinkedIn, occupations and skills are self-reported by members. Not all skills reported by members may be required or needed by employers. Similarly, members may report the skills they think employers look for and under-report skills that they consider less important. This may be the reason why soft skills appear less important in this study than in other recent studies (e.g. Deming and Khan, 2017).

Finally, the results of our analysis illustrate that data reported by members of professional networks and job-search platforms such as LinkedIn can be valuable new sources of information with which to assess rapidly changing skills demands and to increase countries' adjustment capabilities.

However, these alternative data sources complement rather than replace traditional sources of workforce data, such as censuses, household and employer surveys, and administrative records on, for example, education, tax, social security and unemployment. Investment in modern LMI systems is necessary to facilitate the interoperability, sharing and dissemination of different sources and types of data. This intelligence can be shared with a range of stakeholders, including parents and students, workers, employers, policy-makers, and education and training providers, both to generate a more complete and timely picture of the labour market and to facilitate a more rapid and better-informed adjustment of workforce development policies and programmes.

## 3.1.4.  What do new data sources tell us about emerging skills in Latin America and the Caribbean?

Out of the 20 skills that increased the most on average in Argentina, Brazil, Chile and Mexico, ten are directly tied to technological development. There is high growth in the demand for creative digital skills such as game development and animation, digital marketing and computer graphics. Demand for basic digital skills, on the other hand, has fallen, owing to slow growth in many administrative occupations, such as administrative support or accounting, that traditionally required these skills. For the same reason, demand is also lower for other skills associated with administrative occupations.

## 3.1.5.  What options does this research open up for policy-makers and workers?

Users' profile data from LinkedIn can allow us to build tools that help workers and policy-makers answer key questions, such as:

▶ If I want to change occupations, which are most like mine? Which ones are currently growing?

▶ To make the transition to a growing occupation, what skills will I need that I don't have now? And what skills are the most sought after in my geographical area?

▶ Considering my skill set, which occupations now in demand could I move towards?

At the IADB we continues to work with LinkedIn to build such tools. In 2020 we will be working on a career pathways tool that will provide a detailed report for policy-makers wishing to help workers make the transition from declining to emerging occupations in a particular labour market.

# ▶ 3.2. Using the LinkedIn Economic Graph in India

**Sukriti**

## 3.2.1. Introduction to the Economic Graph

With over 645 million professionals as registered members, LinkedIn offers a data set which is global (covering over 100 cities and countries), granular (broken down by location, industry and function), real-time (members are constantly updating their profiles) and historical (showing migration patterns and skills evolution). This dataset forms the Economic Graph at LinkedIn.

The team in charge of the Economic Graph at LinkedIn delivers policy research and labour market insights from the Economic Graph to inform skills, education and employment policy across the world. Our objective is to help the workforce make the transition to the skills and jobs of tomorrow.

We work on four key themes defining the future of work:

▶ **career pathways:** we help policy-makers drive career preparedness and progression among the workforce;

▶ **AI and emerging technologies:** we help economies to take advantage of new technological opportunities;

▶ **entrepreneurship ecosystems:** we provide intelligence about entrepreneurship ecosystems;

▶ **global economic integration:** we help governments connect talent and knowledge to economic growth.

## 3.2.2  Insights from developing/emerging economies on the future of work: India

### Theme 1: Career pathways

We offer policy-makers insights on which jobs employers are hiring for, what skills are in demand, and how people's careers develop.

For example, in 2019 we observed that in terms of volume of hiring, the position of software engineer is top of the list across India's principal industries (ICT, manufacturing and finance). Aside from directly technology-related roles, demand is also high for business management jobs such as business analysts and business development managers.

Technological skills such as SQL, JAVA and C programming language are also in demand, particularly in the software and ICT sectors. Among soft skills, management, team management and leadership are in demand. In terms of career development insights, research undertaken with the WEF recently found that twice as many men as women reached leadership positions at some point during their careers. Moreover, over the course of their careers, women become less and less likely to break into leadership roles.

### Theme 2: AI and emerging technologies

Many "emerging jobs" – that is, jobs advertised with the fastest-growing frequency over the past five years – are related to AI. In the United States, six of the top 15 emerging jobs in 2018 were related to AI: these included blockchain developer, machine-learning engineer/specialist/researcher and data science specialist/manager. The top employers seeking to fill these jobs include Facebook, Amazon, Apple and Google. In India, two out of the top ten emerging jobs are related to AI – these are machine-learning engineers and data scientists, for both of which there is massive demand.

Skills that are increasingly relevant as a result of the rising significance of AI are:

▶ data and programming skills that are complementary to AI;

▶ skills in using products or services that are powered by data, such as search engine optimization for marketers; and

▶ interpersonal skills.

There is a significant gender gap in AI professionals, only 22 per cent of whom globally are female. Moreover, there are no signs that this gap is closing. Men and women have been adding AI skills to their profiles at similar rates, so while women are not falling further behind, they are not catching up either. Women with AI skills are more likely to work in the use and application of AI, while men are more likely to work in the development of the technology itself.

## Theme 3: Entrepreneurship ecosystems

We also notice gender gaps among entrepreneurs, the majority of enterprise founders being men. When we examine the percentage share of companies in a country that are founder-led, we find that India is close behind the United States, with Israel way ahead of both.

## Theme 4: Global economic integration

One out of every three emigrants from India takes up a job in United States, which is the top destination for international talent migration from the country. Other frequently chosen destinations are the United Arab Emirates, Canada, the United Kingdom and Australia. Together, these five countries account for almost 70 per cent of migrants from India. Germany, the United States, the United Kingdom, Australia and Canada (in that order) have the highest in-demand migration rates, i.e. the highest percentage shares of migrants from India taking up jobs that feature among the top ten jobs by volume being hired for in the destination country. One of every five Indians migrating to Germany is taking up an in-demand job such as software engineer, project manager or research assistant.

# ▶ 3.3. Using real-time big data to inform TVET policies and strategies: The case of Myanmar

**Hiromichi Katayama**

## 3.3.1. The data challenge in UNESCO's work to support national TVET policies

### TVET policy process

UNESCO supports its member states in establishing evidence-based policy development and implementation on technical and vocational education and training (TVET). Within the framework of its TVET strategy, UNESCO uses TVET and labour market data to identify and anticipate trends in order to inform member states about the future of skills supply and demand in the labour market. It also supports the development of data-backed policy and programmes.

There are three stages in UNESCO's TVET policy process:

▶ the policy review stage, in which skills supply and demand are analysed and strategic directions for reforms are proposed;

▶ the policy development stage, in which comprehensive skills development strategies are formulated; and

▶ the policy implementation stage, in which a costed plan of the strategies is developed and implemented.

All three stages of the reform process depend on the possession of relevant data.

Owing to the cross-cutting nature of TVET and fragmentation of data/statistics, 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. A central challenge is the lack of data integration. In any country, the Ministry of Labour, the Ministry of Education and the private sector usually hold different data sets relating to TVET and the status of labour supply and demand. While some of this data may overlap, each institution or sector is likely to hold different data based on its own fragmented activities. This data needs to be streamlined and fine-tuned in order to be used effectively. A lack of clear lines of governance to link all the actors and processes involved further inhibits effective data usage. Finally, data that is collected and analysed may not be translated and disseminated to relevant target groups in a systematic way.

## 3.3.2. Supplementing traditional LMI with big data to generate more useful knowledge

### Traditional labour market information

UNESCO's experience in Myanmar demonstrates the potential for combining traditional LMI with big data from online job-search platforms. Traditional LMI offers a detailed picture of the status of Myanmar's labour market. For example, labour force survey data can give us the sectoral breakdown of jobs. Of the 24.1 million jobs registered in 2017, 21.1 million were in traditional sectors, principally agriculture; 1.7 million people were employed in the modern private sector; and a small proportion of workers are in government employment. We can further see employment distribution divided along gender and urban–rural lines. Urban workers are more highly concentrated in trade, hospitality, transportation and utilities, whereas the reverse

is true in agriculture and mining. Females are over-represented in trade, manufacturing and hospitality, while there are more men in agriculture and construction. Traditional data also includes unemployment data, which shows us that both youth and overall female unemployment in particular rose (compared to adult male employment) between 2015 and 2017 (UNESCO, 2019).

Survey data also tells us why employers were having difficulties in hiring to fill vacancies. At the level of managerial, senior professional and semi-skilled non-production workers, there were two key reasons: applicants lacking personal or job specific-skills, and applicants expecting higher wages than could be offered. At the level of skilled or unskilled production workers, the key reason was a lack of either personal or job-specific skills. Information from the directorate of investment and company administration tells us that there will be major infrastructural and industrial development in Myanmar over the next five to ten years, which will fuel job growth. To get the relevant skills, workers will need to go through some combination of the country's fragmented training provision.

## Use of big data to supplement traditional LMI

This picture can be improved with recourse to big data analytics and AI, using data extracted from an online job board. Data was mined from JobNet.com.mm and specific job titles were mapped to appropriate ones within JobKred's taxonomy (see figure 3.1).[54] An initial analysis was then done using the mapped job titles, with frequency of occurrence as the indicator, to arrive at an understanding of employer demand in the labour market. The jobs for which demand was strongest were sales executives, accountants and senior accountants. Subsequently, the job descriptions were processed by JobKred's predictive engine to identify the skills that were relevant to the respective job titles. For example, the top three skills for sales executive

▶ **Figure 3.1. Demand for top 20 job titles in Myanmar in 2019**



Job title
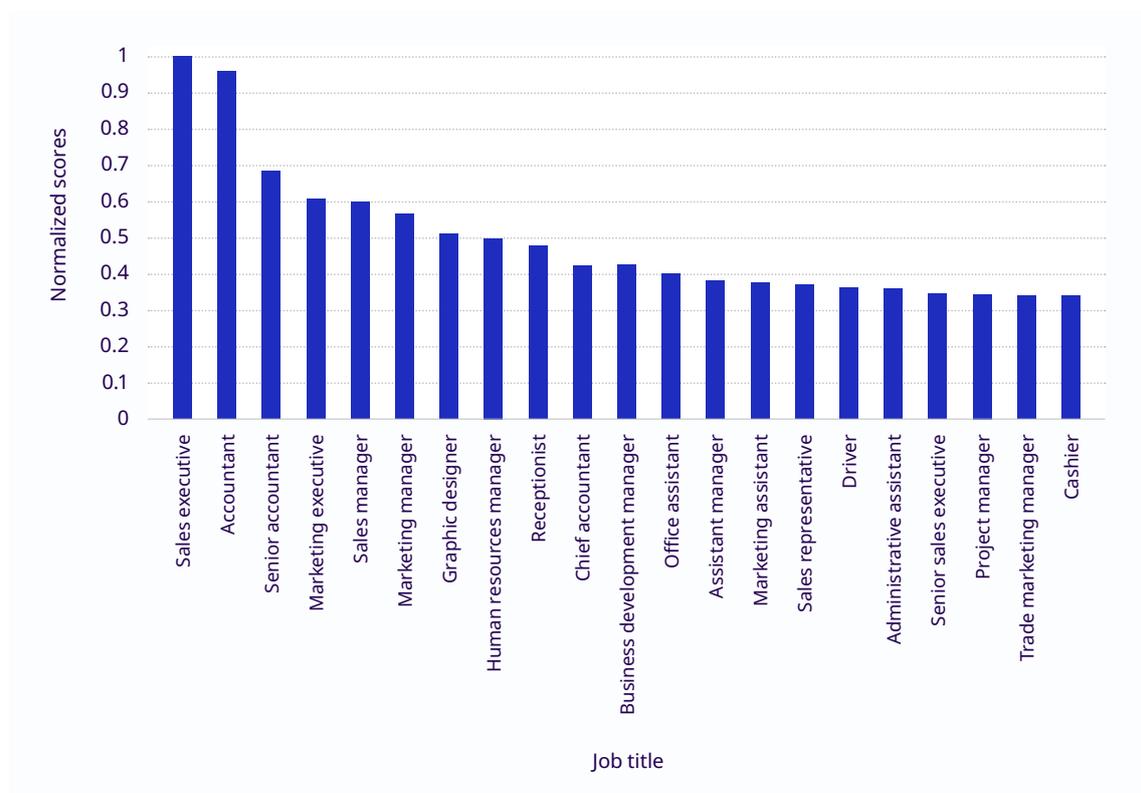
**Source:** UNESCO, 2019, based on data analysis conducted by JobKred (www.jobkred.com).

54 www.jobkred.com [15 Feb. 2020].

positions were sales, identifying new business opportunities and customer service (see figure 3.2). Employer demand for these skills was then calculated, again using frequency of occurrence as the indicator. Overall, the skills most in demand were sales, marketing and identifying new business opportunities. A preliminary analysis was then done with the data obtained from the above research. This data is useful in revealing the status of the labour market and considering improvements to the TVET system.

▶ **Figure 3.2. Skills demand by occupation in Myanmar**

| | Top 10 Skills for sales executive | | Top 10 skills for sales manager | | Top 10 skills for business development manager |
|---|---|---|---|---|---|
| 1 | Sales | 1 | Sales | 1 | Identifying new business opportunities |
| 2 | Identifying new business opportunities | 2 | Sales and marketing | 2 | New business development |
| 3 | Customer service | 3 | Sales management | 3 | Managing client relationships |
| 4 | Managing client relationships | 4 | Sales planning | 4 | Business |
| 5 | Retaining customers | 5 | Managing client relationships | 5 | Business development |
| 6 | Long-term customer relationships | 6 | Sales operations | 6 | Sales |
| 7 | Sales and marketing | 7 | Customer service | 7 | Helping clients succeed |
| 8 | Introducing new products | 8 | Identifying new business opportunities | 8 | Introducing new products |
| 9 | Sales planning | 9 | Retaining customers | 9 | Building relationships |
| 10 | Listening to customers | 10 | Sales growth | 10 | Client services |

**Source:** UNESCO, 2019, based on data analysis conducted by JobKred (www.jobkred.com).

Challenges in the use of real-time data include identification of online job boards and their reliability, the coverage of the data (and issues relating to lack of data, for example, blue-collar jobs or informal employment), data-cleaning problems, taxonomy decisions and complementarity with other LMI (which is necessary to establish holistic skills governance). UNESCO will continue to use big data analytics relating to skills, TVET and the labour market to further support its work around TVET (see UNESCO, 2019).

# 4

# Connecting the dots: Combining big data and other types of data to meet specific analytical demands

This chapter looks into how big data can be used to gain understanding in specific policy contexts and to examine the effects of megatrends on changing skills demand. Examples include the analysis of STEM requirements across different types of jobs and sectors, and understanding and forecasting the effects on jobs and skills demand of technological change, of trade and globalization, and of the implementation of measures to combat climate change. The usability of big data analytics is not straightforward, presenting many challenges and limitations that have to be taken into account in assessing or forecasting demand for skills.
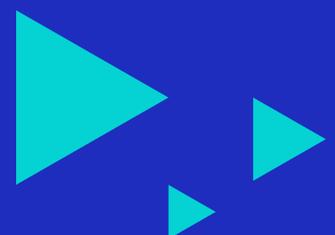
## ▶ 4.1. The STEM requirements of "non-STEM" jobs: Evidence from UK online vacancy postings

**Inna Grinis**

In the United Kingdom, less than half of science, technology, engineering and mathematics (STEM) graduates work in so-called "STEM occupations" (as, for example, scientists or engineers). If, as is often thought, all recruiters in "non-STEM" occupations (for example, graphic designers or economists) neither require nor value science and technology skills, and simply like hiring science graduates for their problem-solving and analytical abilities, this apparent leakage from the "STEM pipeline" should be considered problematic, because a STEM education is more expensive and difficult to acquire than a non-STEM one.

I have shed new light on this issue by developing a novel approach to identifying science and technology jobs through the keywords collected from online vacancy descriptions, rather than, as is typically done, by classifying occupations discretely into STEM or non-STEM, then considering all the jobs belonging to the first group as STEM and the rest as non-STEM.

This approach is made possible by having access to a large data set collected by the firm Burning Glass Technologies, which contains information on all vacancies posted online in the United Kingdom between 2012 and 2016.

I have designed and evaluated machine-learning algorithms to identify STEM jobs as those whose recruiters are more likely to look for STEM rather than non-STEM graduates because the advertised position requires certain skills and knowledge that are taught exclusively, or much more often, in scientific disciplines (for example, "systems engineering"), and/or involves job tasks, tools and technologies for which a STEM education is typically required (e.g. "C++", "design software").

This job-level analysis reveals that STEM jobs should not be equated with STEM occupations. In fact, 35 per cent of all science and technology jobs are to be found in non-STEM occupations, and 15 per cent of all postings in non-STEM occupations are technology jobs.

Moreover, equating jobs with occupations leads to underestimating the overall demand for science and technology skills, since STEM jobs far outnumber jobs in STEM occupations – for example, by half a million employment opportunities in 2015.

I also find that technology jobs are associated with higher wages within both STEM and non-STEM occupations, even after controlling for detailed occupations, education, experience requirements, etc.

Although, overall, my findings suggest that the leakage from the "STEM pipeline" may be less wasteful than is typically thought, because a significant number of recruiters in non-technology occupations do require and value science and technology skills, the issue remains problematic, for two main reasons:

▶ first, nothing prevents STEM-educated jobseekers from taking up non-technology jobs within non-STEM occupations, for which non-STEM graduates are also qualified and no wage premium is offered;

▶ second, by exploring the keywords from the job postings, I found that the technical skills and knowledge posted in STEM vacancies within non-STEM occupations could, in many cases, be acquired with less training than a tertiary science qualification – for example, learning how to code in "C++" does not require a bachelor's degree in computer sciences.

Hence, a more efficient way of satisfying STEM demand within non-STEM occupations could be to teach more science and technology skills in non-STEM disciplines. This education policy could reduce skills shortages in both STEM and non-STEM occupations.

▶ **Figure 4.1.  The geographical locations of STEM vacancies posted in the UK, 2015**

**(a) % of STEM jobs in each county**



▶ London: 14.5%
▶ 4–5%
▶ 3–4%
▶ 2–3%
▶ 2–3%
▶ <1%
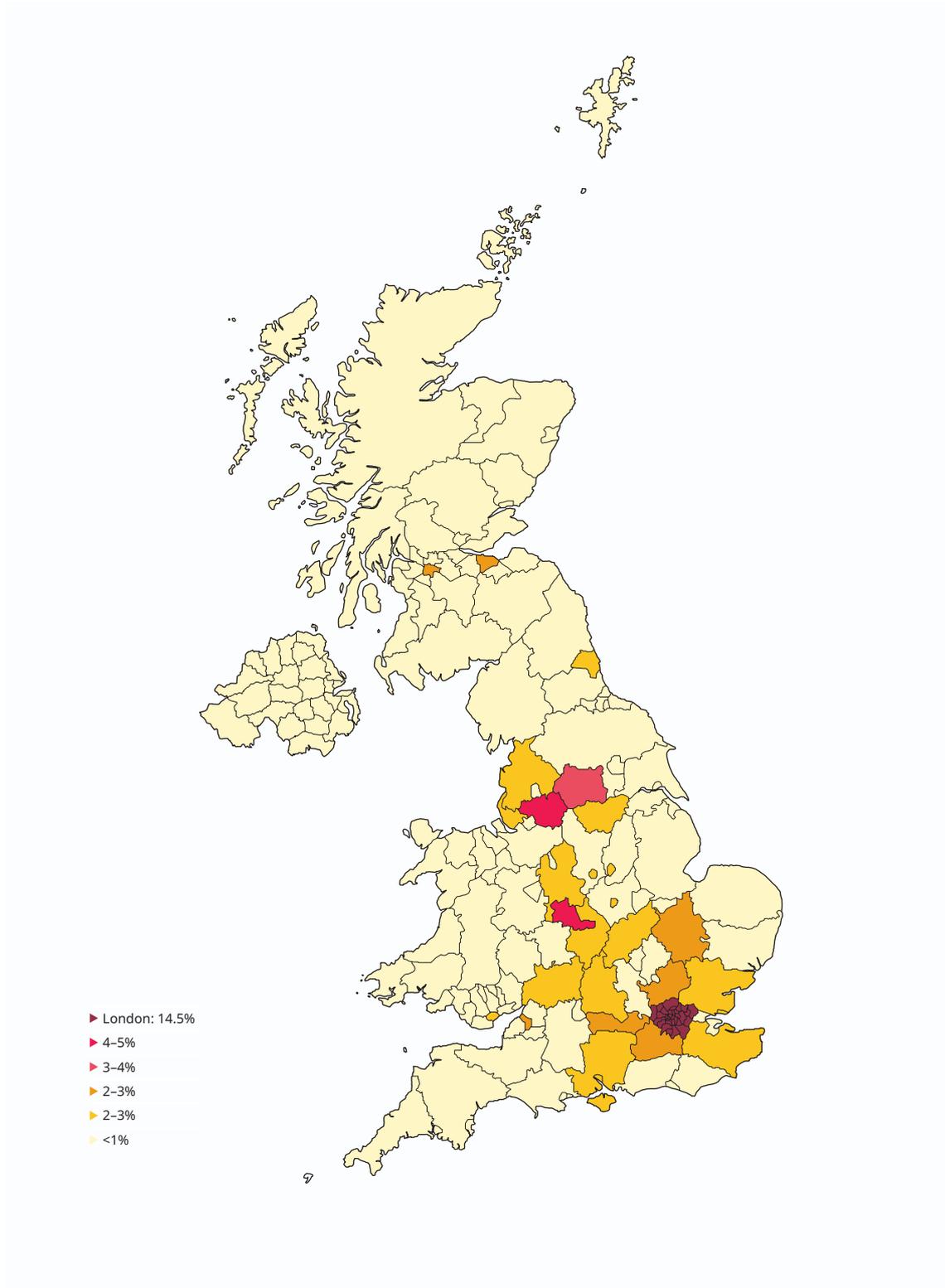
**(b) STEM density of each county**
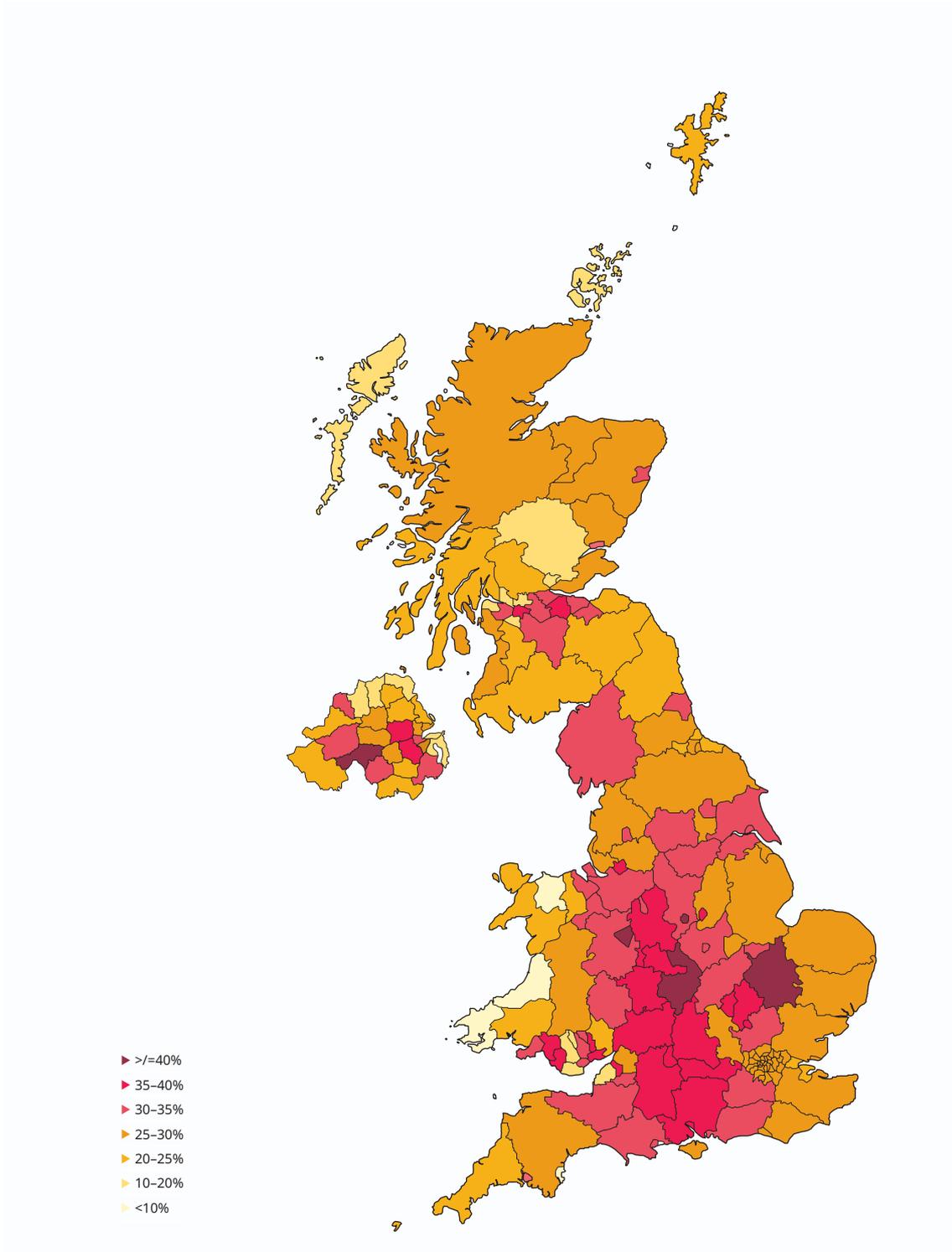


**Notes:** Based on the sample of 77.8% of all vacancies with county identifiers in 2015. London includes the 32 London boroughs and the City of London. STEM density is the percentage of jobs within a county that are classified as STEM. Map (a) is re-weighted using the 2015 Annual survey of hours and earnings produced by the Office of National Statistics (ONS, 2015).

**Source:** Author.

# ▶ 4.2. Using real-time LMI to measure digital transformation [55]

**Julia Nania, Hal Bonella, Dan Restuccia and Bledi Taska**

Digital skills are becoming ever more important in today's economy: employers indicate that in about one-third of cases, difficulty in filling vacancies is, to some degree, attributable to a lack of appropriate digital skills among applicants (Winterbotham et al., 2018). But the term "digital skills" covers a wide array of competencies, knowledge and skills, making it difficult to design interventions to address the need for such skills. Our research, presented in the report *No longer optional: Employer demand for digital skills* (Nania et al., 2019), attempts to illuminate the issue through analysis of millions of online job advertisements in the United Kingdom to highlight the skills employers are demanding. The aim is to provide an overview of digital skills demand and a useful basis for building an evidence-based skills development policy in this area.

The research highlights the importance of both baseline digital skills, such as those required to use productivity software tools, and skills in using more specific software tools that are critical to qualify jobseekers for medium- and high-skill roles. Specific digital skills are key to unlocking opportunities for jobseekers and addressing the shortage of digitally skilled workers in the United Kingdom. Digital skills are essential entry requirements for two-thirds of occupations listed in the UK SOC,[56] and carry with them a wage premium over non-digital roles. These occupations account for 82 per cent of OJVs.[57]

To understand the demand for digital skills at various levels of the job market, we have broken the market down by skill level,[58] distinguishing low-skill jobs that require minimal training, medium-skill jobs that require higher national diplomas/certificates, and high-skill jobs that require a degree or above. We examined the importance of digital skills at each level. Through this analysis, we found that digital skills are becoming near-universal requirements for employment. Progress up the career ladder from low- to high-skill jobs brings an increasing demand for specific digital skills; and acquiring specific digital skills makes both career progression and a pay increase more likely. In certain fields, jobseekers need to develop digital skills related to specific technical tools of their chosen discipline to make any headway in their careers. Examples are computer-aided design for engineers and technicians, search engine optimization for marketers, and data analysis skills such as the programming languages SQL and R for analysts.

We focus here specifically on understanding demand in terms of eight common clusters of digital skills (clusters are groups of skills that are often found together in job postings). The "productivity software" cluster is requested in the vast majority of job adverts across the economy: these are the skills we categorize as "baseline" digital skills. The other seven clusters cover digital skills required within specific roles or sectors: software and programming; networking systems; data analysis; digital marketing; digital design; customer relationship management software; and machining and manufacturing technology. Jobseekers who develop skills in one or more of these clusters can qualify for many of the best-paying and fastest-growing jobs in today's economy.

Key findings from this research include the following:

▶ **Digital skills are near-universal requirements.** "Baseline" digital skills such as the ability to use Microsoft Office and other productivity software tools are commonly required in jobs across all skill levels and have become a ticket to entry into the labour market. When breaking the job market down

---

55  This section is based on the executive summary of the publication *No longer optional: Employer demand for digital skills* (Nania et al., 2019) .

56  The UK SOC classifies jobs on the basis of skill level and skill content: see https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc [15 Feb. 2020].

57  Unless otherwise noted, all data on employer demand for jobs and skills in this report is sourced from Burning Glass's database on online job postings and reflects the 12-month period from 1 Apr. 2017 to 31 Mar. 2018.

58  Definitions on skill levels used in this report can be found in the "Methodology" section of Nania et al., 2019.

by skill level into low-, medium- and high-skill roles, we find that over 75 per cent of job openings at each level request digital skills.

▶ **Digital skills carry a wage differential.** Overall, roles that require digital skills pay 29 per cent (£8,300 per annum) more than roles that do not (£37,000 p.a. vs £28,700 p.a.). This difference is apparent at all skill levels, but the differential increases at higher levels. The salary differential for digital skills ranges from £2,700 for low-skill jobs (£24,000 vs £21,300) to £5,800 for medium-skill jobs (£32,200 vs £26,400) and £11,300 for high-skill jobs (£45,300 vs £34,000).

▶ **Digital skills are in demand everywhere.** Digital skills are required in at least 82 per cent of OJVs across the United Kingdom, but the precise skills demanded are not uniform across the country. For example, the capital region has the greatest demand (digital skills required in 87 per cent of advertised roles, spread across almost all sectors), while the west midlands has a slightly lower demand at 82 per cent of roles; but the prominence of the manufacturing sector in the latter region means that machining and engineering software skills are required in 24 per cent of those roles.

▶ **Specific digital skills may help workers avoid the risk of automation.** By entering a role that requires specific digital skills, workers can reduce their risk of being displaced by automation by a dramatic 59 per cent. Specific digital skills commonly complement uniquely human skills such as design, writing or communication, which in combination are difficult to automate and critical to a firm's success.

▶ **Specific digital skills promote career progression.** To maximize chances of success in the digital economy, jobseekers must go beyond baseline digital skills and develop more specific skills. Importantly, these specific digital skills are not required only in the technology sector but are in demand across all sectors of the economy. These may include the ability to use digital tools such as Adobe Photoshop for designers; computer-aided design for engineers and manufacturing workers; customer relationship management software for sales and marketing professionals; and computer programming and networking for IT professionals. These specific digital skills are required in 28 per cent of low-skill jobs, 56 per cent of medium-skill jobs and 68 per cent of high-skill jobs.

Implications of this research for future policy development include the following:

▶ **Jobseekers need a complete package of skills, both digital and non-digital, to succeed in the economy.** Many of the specific digital skills covered in our research serve to enable non-digital expertise. For example, software programs such as Adobe Photoshop serve to enable design work, and customer relationship management software tools can make sales and marketing professionals more effective in communicating the messages they craft.

▶ **Digital skills policy should be driven locally.** Digital skill requirements vary substantially from region to region, and so should efforts to train workers. For example, data and design skills are particularly important in London to meet the needs of the finance and creative industries, while engineering and advanced manufacturing skills are particularly important in high-tech engineering centres such as Cambridge and Bristol. In Northern Ireland, demand for baseline digital skills and lower-paying roles accounts for a disproportionate share of demand compared to specific digital skills. Because of these different needs, each region should aim to develop a digital skills policy that matches the demand of local industries and develops the skills that may be needed to execute an economic development policy based on human capital.

▶ **Digital skills will change over time**. Policy should be shaped to anticipate dynamic development in the digital economy. Many of the fastest-growing skills and software packages in areas such as data analysis and digital marketing hardly existed just a few years ago. Skills policy must accordingly be dynamic, continually re-evaluating the skills that are and will be in demand to train workers for the roles of the future, while also ensuring they have the skills needed for success today. Such an approach by employers allows them to build and shape their internal workforce in response to changing market and business needs. A dynamic approach to skills evaluation by educators consists of constantly re-evaluating curriculum content to ensure that students have the particular sets of skills they need to get jobs in their region.

# ▶ 4.3. Sharing experiences in using big data in combination with other methods

**Ana Podjanin, Olga Strietska-Ilina and Bolormaa Tumurchudur-Klok**

This section focuses on work done by the ILO in relation to the use of big data, or more precisely OJV data, for skills needs anticipation and analysis. The first subsection looks at some of the main challenges related to the analysis of vacancy data from developing countries. The rest of the section presents some of the most recent ILO efforts in combining big data analysis with traditional data sources, and discusses the limitations and advantages.

## 4.3.1. Challenges for developing countries and beyond

Skills imbalances appear to be a global problem, although manifested across regions and countries in very different ways. While in many high- and middle-income countries a problem of overqualification often prevails, in lower-income countries underqualification is a predominant problem. At the same time, we also know that even in low-income countries, where access to training is problematic, there is still a large proportion of overqualified people. This is often caused largely by a deficit of decent jobs offering working conditions and wages attractive enough to absorb the talent available on the labour market. But also it is partially caused by deficiencies in specific skills and competencies required on the labour market, where a person may be overqualified but underskilled. Identifying the particular skills and competencies that are needed on the labour market is therefore a key element in the labour market information system (LMIS). It can inform the design of competency standards and training programmes in both advanced and developing countries, and thereby ensure the relevance of education and training provision to the needs of the labour market.

However, information on specific skills and competency needs is often a missing link in the LMIS. Usually, such information is collected by qualitative and semi-qualitative methods, such as occupational research, expert committees, or surveys among employers and workers. Such approaches may be rather costly, especially if the information collected is to provide the level of granularity that is needed as a basis for analysis that can inform the design of education and training. Conversely, the use of quantitative methods presents a series of challenges, especially in developing countries. The quantitative measurement of skills and competency requirements in a comparable and robust way requires the use of proxies, such as occupations or levels of educational attainment; but such proxies do not provide the necessary level of detail to inform the design of training programmes. In addition, in developing countries, standard sources of statistical information, such as labour force surveys (LFS), may not be updated regularly and their results may not be very accurate, owing to, among other things a lack of sufficient capacity in national statistical systems. This lack of capacity is often mainly attributable to a lack of resources, as carrying out regular and high-quality surveys is a very costly exercise.

The use of real-time big data analytics may offer a way to resolve these problems. As noted in previous chapters of this report, OJVs constitute a very rich source of information for skills needs identification. They can be especially powerful and effective if complemented by more traditional methods. However, like most data sources, they also have their limitations, in particular when it comes to developing countries. Access to the Internet and levels of literacy play an important role when it comes to accessing online job-search tools. Furthermore, in developing countries where the informal economy is very large, OJVs cover only a small proportion of actual vacancies, as it is very likely that only the formal sector is represented in online job platforms. Similarly – and this applies to both advanced and developing economies – low-skilled jobs tend to be less well represented among OJVs. These factors emphasize once again the presence of significant challenges when using OJV data – many of which loom large in advanced countries, let alone developing

ones. Therefore, where the focus is particularly on developing countries, it is important to keep these limitations in mind and consider data analysis with strict attention to contextual circumstances.

## 4.3.2. Complementing different data sources: Skills for a greener future

This subsection describes how qualitative country studies were combined with quantitative modelling and big data analytics to produce the report *Skills for a greener future* (ILO, 2019), in order to incorporate skills needs analysis and an occupational outlook into the transition towards low-carbon economies and environmental sustainability .[59]

Qualitative analysis, which covered 32 countries, examined environment-related policies and their coherence with skills development policies, their effects on economic activities and related skills, and good practices in skills development measures. The analysis was based on in-depth country studies and a small-scale expert survey. These elements were complemented by quantitative analysis, based on the multiregional input–output model EXIOBASE v3 and labour force survey data, applied to transactions in 163 industries across 44 countries in order to quantify the occupational skills needs in two scenarios, namely energy sustainability and a circular economy.[60] The model explores the likely job impacts by 2030 of keeping the rise in global temperature below the 2°C ceiling set by the Paris Agreement on Climate Change in 2015. By weighting the results to reflect employment composition in a range of countries, global scenarios were produced (ILO, 2019).

This process enabled the researchers to see which occupations might benefit from employment growth and where some shrinking of employment might be expected. This was a very important quantitative addition to the qualitative research, offering a global coverage and, among other things, permitting a better understanding of potential gender effects by skill level under the scenarios examined. However, as noted above, occupations and skill levels are insufficient proxies for measuring skills and competency needs. Therefore, real-time big data (OJVs) from Burning Glass Technologies (BGT) on tasks and skills required in job vacancies were incorporated into the modelled results by occupation. The limitation of this approach lay in the use of US OJVs data as a proxy in the absence of similar data for all countries; however, no matter how imperfect the approach, it allowed a much more granular understanding of the skills effects in the two scenarios than would otherwise have been possible.

To illustrate the approach, below we provide an example of the occupational outlook for the scenario of a transition to clean energy (energy sustainability). It is estimated that the global impact on employment of a transition to energy sustainability by 2030 will entail the creation of almost 25 million jobs and the loss of nearly 7 million jobs. Of the latter, 5 million can be reclaimed through labour reallocation – that is, 5 million workers will be able to find jobs in the same occupations in another industry within the same country. The remaining 2 million workers are likely to be in occupations that are not reallocatable – that is, where jobs will be lost without equivalent vacancies arising in other industries (ILO, 2019).

Some occupations, will experience high levels of job creation with little or no job destruction; others will undergo high turnover, with both job creation and job destruction, requiring reallocation of workers across industries with some retraining to adapt workers' skills to the new industry context. Figure 4.2 shows occupations that will experience net job creation or loss, ranked by the highest levels of job reallocation (ILO, 2019).

The creative destruction of jobs will entail the reallocation of almost 750,000 science and engineering associate professionals (ISCO-31),[61] around 500,000 science and engineering professionals (ISCO-21), and over 300,000 each of stationary plant and machine operators (ISCO-81), drivers and mobile plant operators (ISCO-83), and electrical and electronic trades workers (ISCO-74) (ILO, 2019).

---

59  This subsection is based on the findings of the report *Skills for a green future* (ILO, 2019).

60  Earlier results of the EXIOBASE v3 model were used in ILO, 2018, to produce industry-level estimates. Incorporating LFS data in these results an occupational outlook to be produced (ILO, 2019). A detailed explanation of the model can be found in ILO, 2019, box 6.1.

61  All occupational categories listed refer to the ISCO-08 classification

▶ **Figure 4.2. Jobs created and destroyed in the energy transition scenario by occupation, to 2030: Occupations with the highest reallocation of jobs across industries**
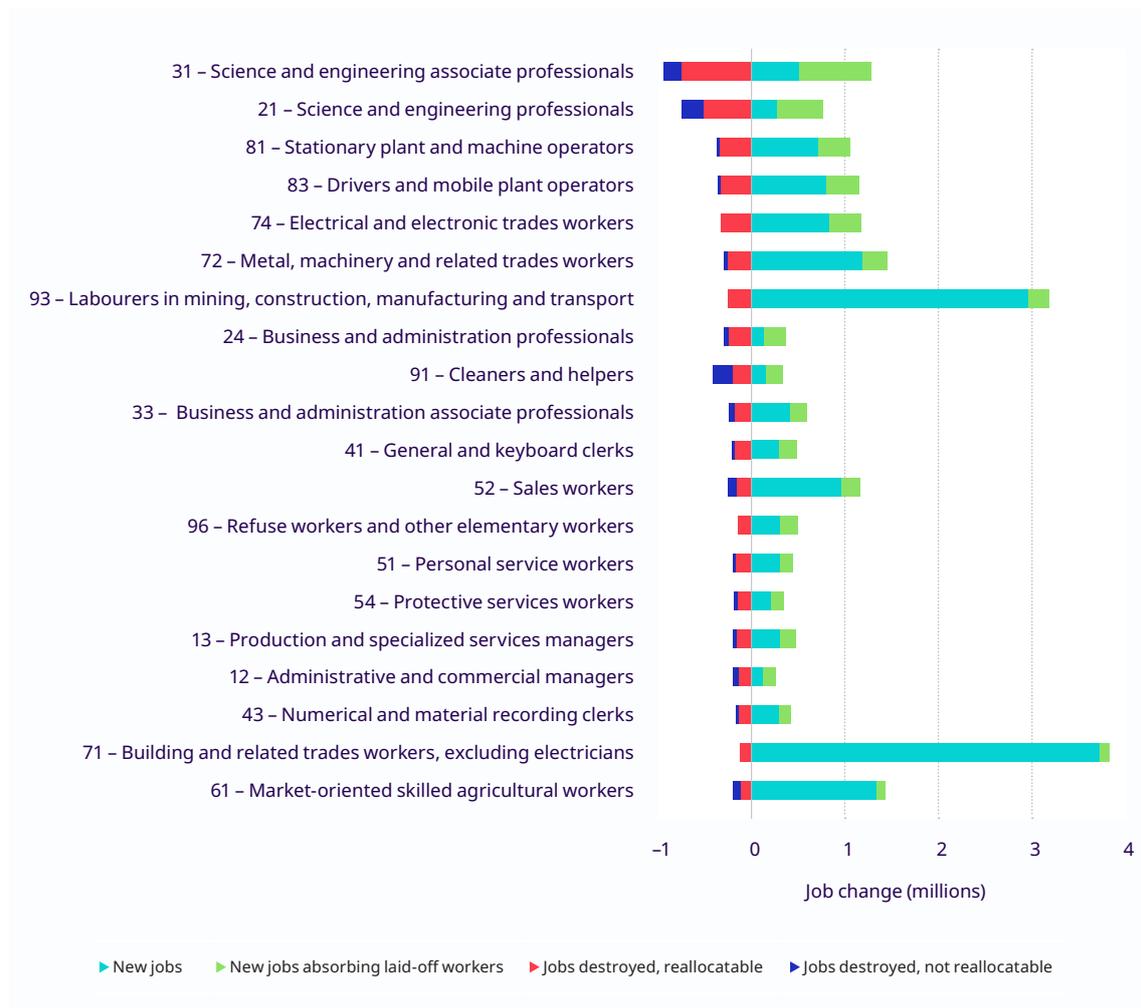


**Note:** Occupations measured at the ISCO-08 two-digit level. The figure shows the 20 occupations with the highest level of new jobs absorbing laid-off workers. "New jobs absorbing laid-off workers" are jobs that can be filled by similar (reallocatable) jobs lost in other industries in the same country or region ("Jobs destroyed, reallocatable"). "New jobs" are jobs created that cannot be filled by jobs lost in similar occupations within other industries in the same country or region. "Jobs destroyed, not reallocatable" are jobs for which vacancies in the same occupations in other industries within the same country or region will not be found.
**Source:** ILO, 2019.

Some viable transition paths for workers losing their jobs in shrinking industries were calculated on the basis of real-time big data on scraped US job advertisements (BGT data). Of course, such an approach has its limitations and should be treated with caution, as an illustration of possible job paths rather than as career or policy guidance. The main underlying assumption is that those who currently hold jobs that require specific skills and knowledge typically possess relevant skills and knowledge. The viability of a job transition was calculated on the basis of scores of similarity between the job requirements of any given two jobs in terms of the overlap between the activities or tasks that need to be performed. The similarity scores also took into account wage continuity before and after the job transition, the ideal being that the new job is at least as well paid as the previous one (ILO, 2019).

In the energy sustainability scenario, science and engineering associate professionals (ISCO-31) will experience the largest reallocation of workers across industries within this broadly defined occupation. The

results of illustrative calculations of similarity scores for related occupations show that there will be many options for transition to new jobs under this scenario. For instance, power plant operators (unit group 3131 of sub-major group science and engineering associate professionals – ISCO 31) will be able to find jobs as gas pumping station operators (owing to the shift to natural gas), water and waste-water treatment plant system operators and electro-mechanical engineering technicians in other areas (figure 4.3). Some reskilling and upskilling may be needed, especially in relation to the specific machinery and technologies used in new jobs.

▶ **Figure 4.3.  Transition paths for power plant operators (ISCO 3131) under the energy sustainability scenario**

**Power plant operators**
Power production
plant operators

**5**

**opportunities with the same salary or higher, out of total 139 opportunities with high and medium similar scores**

**Gas compressor and gas pumping station operators**
Petroleum and natural gas refining plant operators
Similarity scores: 0.955

**Water and wastewater treatment plant and system operator**
Incinerator and water treatment plant operators
Similarity scores: 0.928

**Mates – ship, boat, and barge**
Ships' deck officers and pilots
Similarity scores: 0.905

**Non-destructive testing specialist**
Mechanical engineering technicians
Similarity scores: 0.88

**Electro-mechanical technicians**
Electrical engineering technicians
Similarity scores: 0.854

**Note:** The calculations are based on a similarity score methodology elaborated by BGT for the WEF (WEF and BCG, 2019). The score takes into account possible job transitions based on a similarity of requirements of two jobs and overlap of tasks, skills and knowledge, education and years of work experience without wage losses.
**Source:** ILO, 2019.

In order to know how to adequately reskill workers who are currently in jobs expected to be destroyed in the transition to greener economies, it is important to identify the associated skills sets in the new jobs forecast to be available. As well as the particular need to pinpoint the skills that individuals should seek to acquire to qualify for emerging jobs in growing industries, it is also interesting to capture the skills required for the jobs that are expected to shrink, in order to identify which skills overlap and are therefore common, i.e. transferable between jobs.

In identifying the jobs that will be reallocatable between declining and growing industries within the same occupations, it is assumed that the core sets of both technical and soft skills and knowledge will be re-used, as they remain the same in principle within each occupation. The real-time big data for the United States in 2017 (BGT data sets) used as a proxy allowed the researchers to take a closer look at the disaggregated level for the occupation of science and engineering associate professionals (ISCO 31), which tops the list of occupations where jobs can be reallocated across industries. The analysis of skills demanded in job advertisements, related to the industries expected to decline and to grow, is presented in the word cloud

▶ **Figure 4.4. Overlap of skills for science and engineering associate professionals in declining and growing industries (energy sustainability scenario)**

# Energy sustainability

### Lower demand          Higher demand

Seismic
inversion
Silica
Energy trading
Geologic data
interpretation
Geological maps
Benzene
Infor software
Snubbing
Instrument design
Structural geology
Transportation security
Hydraulic fracturing
**Petrel softwar**
**Geosteering**
**Petrophysics**
**Wireline**
**Nuclear plant design**
**Geophysics**
**Well testing**
**Reservoir simulations**
**and evaluations**
**Gas prospecting**
**Knowledge of meridium**
**system**
**Radiation protection**
**Sales**
Torque/Drag analysis
Well stimulation
Commercial
development
Facies

Knowledge
of retail industry
Repair
Knowledge of calculus principles
Graphic and visual design software
Quality assurance and control
**Product development**
**Organizational skills**
**Writing**
**Drafting and engineering design**
**Mechanical engineering**
**Creativity**
**Planning**
**Customer handling**
## Problem solving
## Communication
## Microsoft Office
## Project management
**Teamwork / Collaboration**
**Research**
**Scheduling**
**Budgeting**
**Sales and marketing skills**
**Troubleshooting**
Computer literacy
Attention to detail
**Manufacturing processes**
Technical support
Construction management
Multitasking
Industrial engineering industry
expertise

Mockups**
Mechanical, Electrical,
and Plumbing (MEP) Design
Product location
Requirements analysis
Architectural engineering
Plumbing design
Rapid prototyping
Material flow management
Plumbing systems
**Commercial construction**
**Store management**
**Store merchandising**
## Design-build
## Vehicle systems
## Shop drawings
## Connected home
## Design for Six Sigma
## Lead follow-up
**Infotainment**
**Automotive engineering**
## Mechatronics
Schematic design
Project architecture
Catia V5
PCB layout and design
Vehicle design
Tool design
UX Wireframes
Process validation
Site surveys

**Note:** Skills are ranked on the basis of the frequency with which they are requested within the occupational category by growing or declining industries, with those most requested appearing larger. The dark blue area in the middle shows a large overlap of skills within the same occupations in both declining and growing industries. US data (BGT) are used as a proxy.   * also called "soft news" – type of media combining information and entertainment.   ** In manufacturing and design, a mockup, or mock-up, is a scale or full-size model of a design or device, used for teaching, demonstration, design evaluation, promotion, and other purposes

**Source:** ILO, 2019.

(figure 4.4), based on the frequency with which the skills are requested within the occupational category (those most requested appear larger). Industries have been grouped into those expected to grow and those expected to decline. In the middle, an overlap of skills required in both declining and growing industries appears. The analysis revealed that there will be three types of skills that can be re-used and that therefore constitute the core employability skills potentially useful for securing new jobs: (i) soft skills, both cognitive and non-cognitive, such as communication, problem solving, customer handling, teamwork and

organizational skills; (ii) semi-technical transferable skills, such as ICT, scheduling, sales and marketing, budgeting and planning; and (iii) technical transferable skills, such as repair and preventive maintenance (ILO, 2019).

The big data analysis allowed the researchers to understand the composition of transferable skills at a granular level in relation to occupational category. It revealed that it is not only soft skills that are transferable, but also technical and semi-technical skills which may be occupation-specific. The important point here is that "occupation-specific" does not imply "job-specific", as these skills will be important for securing jobs in a range of growing industries. This observation has important implications for TVET policies, pointing to the need to adjust the initial training of the future workforce in such a way that a good set of soft, semi-technical and transferable technical skills are at the core of curricula and competency standards (ILO, 2019).

## 4.3.3.  Validating and complementing results of qualitative sectoral studies on Skills for Trade and Economic Diversification (STED)

This subsection discusses the ILO's effort to use OJV data for the United States (again from BGT) as a proxy for understanding and anticipating skills priorities in responding to change in skills demand in the context of global trade. In particular, OJV data for the US states of Florida, Indiana and Kentucky and Michigan was used to corroborate the findings and results of the ILO's work in developing countries on skills required for growth and development and for resilience in the face of change.[62]

The ILO's STED programme supports countries and sectors in identifying the skills needed for success in trade, and in applying a systemic approach to designing and implementing strategies to address the skills constraints and gaps identified. The programme has been implemented in more than 30 sectors and in around 20 countries. Examples of skills needed to ensure that vital business activities can be implemented effectively and where gaps commonly exist include marketing, science and engineering for product/process development and improvement, industrial engineering, human resource management, and supply chain management. These all point to the urgent need to respond not only to the requirements of trade but also to those of other drivers such as technological change, digitalization, climate change and change in work organization. These labour market trends, as well as  the need for compliance with regulatory requirements and standards, impose new and changing skill requirements. As a consequence, effective skills development systems must be responsive and should include both core employability skills, including transferable skills, and occupation-specific technical skills.

Using OJV big data from BGT on skills and tasks advertised as being required for jobs along with codes for industry and location, it is possible to gain important insights relevant to skills and trade policy in industrialized countries that suffer from employment shocks in tradable sectors. The analysis also serves to corroborate conclusions reached for tradable sectors in developing countries on the basis of findings from STED work. We look here at the case of the United States, which experienced a regional employment shock owing to changes in trade and technology, with the aim of understanding the skills needed during the recovery of manufacturing employment that has occurred in some states.

Manufacturing employment was mostly flat in the United States in the 1990s, with some regional variations, but in the decade after 2000 33 per cent of manufacturing employment was lost, with major episodes of job losses during the economic downturns of the early 2000s and the global economic crisis of 2009. Loss of employment was especially severe in the state of Michigan, where almost half of all manufacturing employment was lost. Across the whole country, there has been some recovery in manufacturing employment since 2009, although the share of manufacturing in total employment continues to fall. However, some of the US states that experienced severe loss of manufacturing employment during the 2000s have experienced a significant recovery since the low of 2009. These include Florida, Indiana, Kentucky and Michigan (see figure 4.5).

---

62  This work has been done in the framework of the ILO's *Global Employment Policy Review* (ILO, forthcoming 2020).

▶ **Figure 4.5. Manufacturing employment in four US states and in all US: Change between 2000 and 2018 (%; 2000 = 100%)**
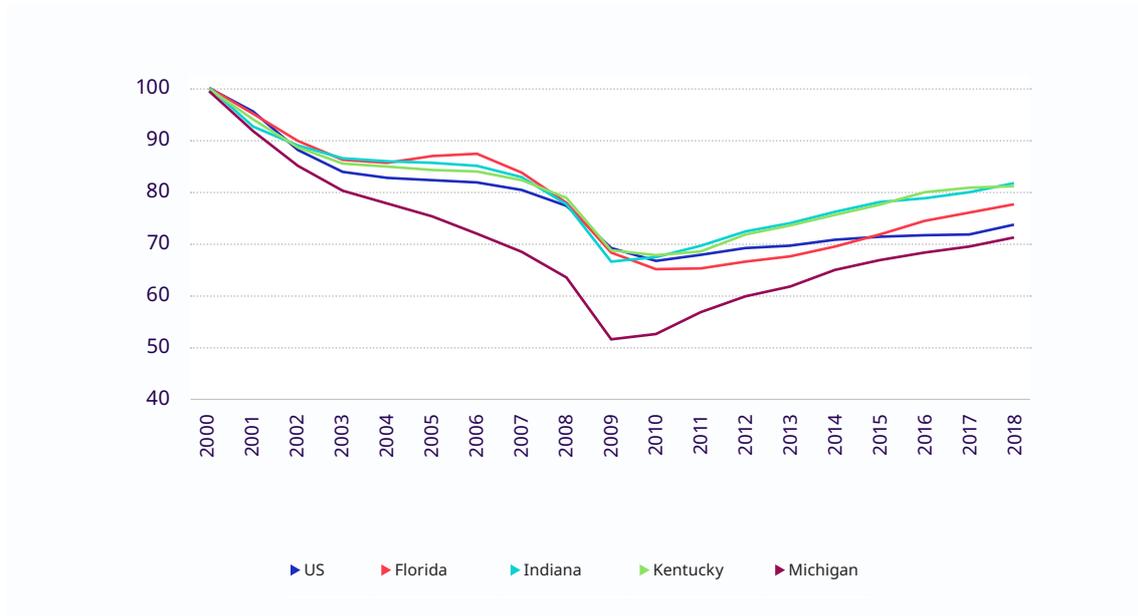


**Source:** ILO, forthcoming 2020.

Focusing on the selected four US states, the following questions were considered: First, how has the composition of demand for skills in the manufacturing industry changed as the upward trend in employment has been sustained? Second, to what extent does the change in demand for skills in these states, which have experienced severe job loss followed by a significant recovery in employment, differ from the experience of the United States as a whole?

BGT data sets indicate the specific skills advertised as being required in each job, where these are specified in the advertisement, and also code for indicators relevant to skills analysis including occupation, any qualifications requirements specified, minimum experience, industry and location. In our analysis, incidence of a skill in job advertisements is calculated as the share of manufacturing industry job advertisements in the occupational classification in which the skill is mentioned, excluding advertisements in which no specific skills are mentioned. The increase is calculated as the incidence in 2015–17 minus the incidence in 2011–13, with the incidence of vacancies analysed at state level. Three-year averages are used to smooth volatility in the data. Ranking is based on a simple mean of the increases in incidence across the four selected states. Thresholds are applied to determine whether the increase in incidence of vacancies for a skill is sufficiently consistent across the four states to be included in the analysis: these are that the coefficient of variation should be less than 1, and that there should be advertisements for the skill–occupation combination in all four states in 2017.

The results highlight themes including digital skills, core employability skills and modern work organization, and transferable skills.[63]

## Digital skills

Advertisements for jobs at all levels have become more likely to specify a requirement for IT user skills. At higher occupational levels this appears as an increasing incidence of the requirement for skills in Microsoft Office, and at clerical level both in Microsoft Office and in spreadsheets. At medium to lower occupational

---

63 The findings presented here draw significantly on chapter 2 of ILO, forthcoming 2020.

level, it appears as an increasing incidence of the requirement for computer literacy or for skills in data entry. At several occupational levels, the increase in the incidence of demand for IT user skills for manufacturing in the four targeted states is significantly above that for manufacturing in the United States as a whole. The share of job advertisements requiring skills in the use and application of more specialized IT systems has increased in general. A need for skills in enterprise resource planning (ERP) correlates with a need for skills in ERP systems. An increasing demand is visible for professional-level workers with skills in graphic and visual design software and in Siemens Teamcentre, and at technician level for skills in SolidWorks. Among advertisements for jobs at professional level, the analysis shows a growth in incidence of requirements for skills in big data and in robotics, both increasing by more than the average for US manufacturing. The underlying data shows that the requirement for skills in big data increased steeply in 2017, the latest year covered by the analysis.

## Core employability skills and modern work organization

The incidence of core employability skills in job advertisements in the manufacturing sector has increased for several occupational categories. For ISCO 8 (plant and machine operators), communication skills is now ranked first, having increased at a rate significantly above the US average for this occupational category for the four case study states. Customer handling, which relies highly on interpersonal skills, has increased in incidence in advertisements for professionals, for clerical support workers, and for craft and related trades workers. Core behavioural skills, such as orientation to detail or positive disposition, appear for a number of occupational categories. An increased focus on skills in modern forms of work organization and manufacturing practices also features. Several of these appear prominently in the analysis of manager-level advertisements, including quality management, ERP, quality assurance and control, advanced product quality planning, lean manufacturing and 5S methodology. Several others also appear in other occupational groups, for example root cause analysis, preventive maintenance, work area maintenance, troubleshooting and direct store delivery.

## Transferable skills

Some of the skills highlighted are among the set of skills variously termed generic skills, soft skills or core employability skills. Some are specific technical skills. Some relate to knowledge of specific manufacturing industries that are significant across the four states considered. Some relate to specific administrative roles that are common across most or all businesses. However, most of the skills highlighted in the analysis lie in a middle ground between these categories, being relevant across a range of detailed occupations and manufacturing industries. These skills are transferable between occupations, giving them a particularly urgent relevance to workers at risk from shocks to employment. Such shocks can be driven by structural change, including trade-led and technology-led employment shifts. The analysis provides insight into a broader range of skills likely to be transferable and in demand where there is a prospect of some recovery in manufacturing employment following a negative employment shock. Transferable skills are more broadly relevant to the career resilience of workers in the context of the changing future of work.
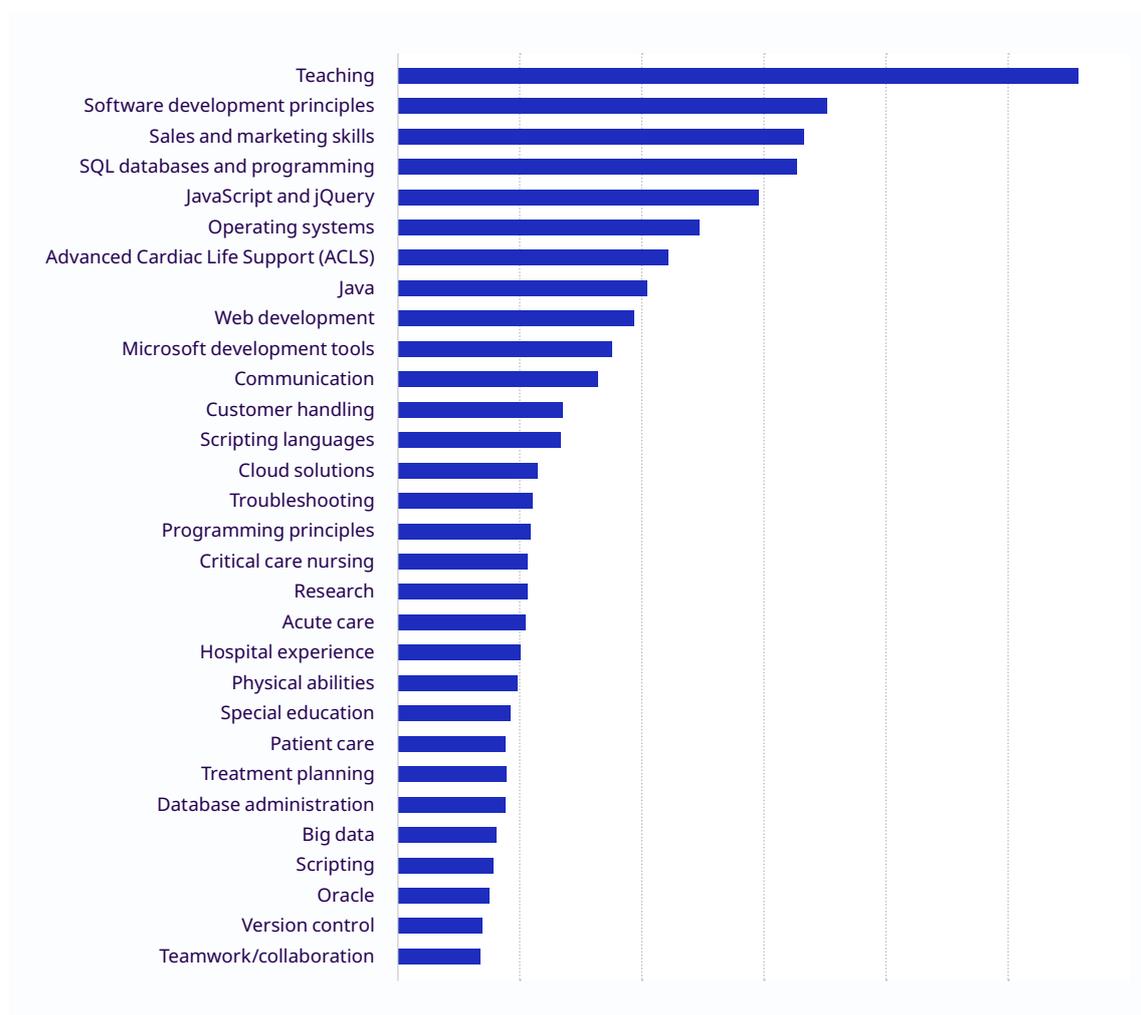
The findings summarized above confirm and consolidate the results and evidence of the ILO's policy work on skills priorities in responding to change in skills demand in developing countries through its STED programme.

## 4.3.4.  Adding granularity and "realtimeliness" by combining labour force survey data and big data

In 2019, the ILO Skills Department embarked on a partnership with the Organisation for Economic Co-operation and Development (OECD) with the purpose of expanding the already established Skills for Jobs Database,[64] and thereby enabling a cross-country comparison of skills for which demand was higher ("hard

---

64  The OECD Skills for Jobs database may be consulted here: https://www.oecdskillsforjobsdatabase.org [26 Apr. 2020].

▶ **Figure 4.6.  Top 30 skills in shortage, related to high-skilled occupations, Uruguay, 2017**

| Skill |
|-------|
| Teaching |
| Software development principles |
| Sales and marketing skills |
| SQL databases and programming |
| JavaScript and jQuery |
| Operating systems |
| Advanced Cardiac Life Support (ACLS) |
| Java |
| Web development |
| Microsoft development tools |
| Communication |
| Customer handling |
| Scripting languages |
| Cloud solutions |
| Troubleshooting |
| Programming principles |
| Critical care nursing |
| Research |
| Acute care |
| Hospital experience |
| Physical abilities |
| Special education |
| Patient care |
| Treatment planning |
| Database administration |
| Big data |
| Scripting |
| Oracle |
| Version control |
| Teamwork/collaboration |

**Source:** Author's calculations based on Uruguayan Household Survey (Encuesta Continua de Hogares) and US online vacancy data from BGT (reference year 2017).

to find/in shortage") and lower ("in excess") across national, and if feasible also regional, labour markets. While the OECD participating team has established and developed this database for a selection of advanced economies and a few emerging economies, the ILO team aims to expand the geographical coverage to a selection of developing countries across the globe. As outlined in the methodological note developed by the OECD (OECD, 2017), the skills for jobs indicator is produced by using labour force survey data to identify the occupational categories where demand on the labour market is lesser or greater. This indicator will be referred to as the "occupational shortage index". Eventually, these results will be integrated with the skills requirements associated with each occupational category, based on the O*NET (Occupational Information Network) skills taxonomy for the United States. The use of O*NET constitutes a significant limitation in respect of identifying skills needs in specific national contexts outside the United States; however, given that the O*NET taxonomy is constructed in a way that allows statistical analysis, by indicating the degree of importance of each skill for every single occupational category, among other dimensions, and given its comprehensive structure, it is still considered for the moment as the best option.

A general limitation of national skills taxonomies, affecting not only O*NET but any other existing skills mapping, is its rigidity over time. Updating such a classification is very time-consuming and usually very

expensive. Therefore, the responsible team at the ILO team started developing an alternative approach on an experimental basis. This entailed using, instead of the O*NET taxonomy, an "alternative mapping", generated on the basis of the BGT vacancy data set for the United States. The idea behind this is to use real-time information derived from OJVs to identify which are the skills most in demand across occupations. Information in skills demand derived on the basis of the frequency with which skills occur in OJVs is being combined with the previously developed "occupational shortage index" to identify those skills that seem to be harder to find and those less frequently requested in job postings. Nonetheless, the results derived from this analysis should be treated with caution. As noted in previous chapters of this report, using information on skills requirements derived from vacancy data may be associated with significant issues of representativeness as well as omitting so-called "implicit skills", i.e. skills that are not explicitly listed but are implicitly required for the job.

In order to illustrate this experimental approach, an analysis for Uruguay is presented here, based on its household survey data.[65] As OJV data for Uruguay was not accessible to the ILO team at the moment of the publication of these results, for comparability with the results obtained through O*NET the BGT vacancy data for the United States was used to identify skills requirements across occupations. The idea behind this is in the first instance to develop a methodology which might then be applied at a later stage by using OJV data from the target country, in this case Uruguay. In presenting the results, occupational categories were divided according to broad ISCO skill levels (ILO, 2012), further clustered into three broad categories, namely high-, medium- and low-skilled occupations.

As may be seen from figure 4.6, taking Uruguay as an illustrative example, the skills most requested across OJVs seem to be in line with expected results, indicating a clear potential in using such an approach for the analysis of skills requirements. Nonetheless, it is important to keep in mind that these are preliminary results, and that further assessments need to be done before such results may be used for policy-related or any other purposes beyond simple statistical analysis.

## 4.3.5. Conclusion

As illustrated through the examples presented in this chapter, OJV data can be used for a number of different purposes, besides the mainstream identification of skills needs. As shown above, combining this relatively new source of information with more traditional methods and data sources, and applying it to specific contexts, can shed new light on specific areas of study. This type of data can also complement and verify some findings achieved by qualitative research, and fill in information gaps for developing countries. There is still much important work to be done by everyone working with these sources in order to identify how to make best use of this rich bank of information derived from OJVs. Further work will be needed to derive big data from real-time OJVs in middle- and low-income countries to make the findings more specific and relevant to them. Also, innovations in combining big data with traditional methods and sources do not have to be limited to OJV data; further work may well include administrative records, social security registries and other types of information. Creativity in approaches will have to be combined with some experimentation in order to better understand the risks and limitations attending the use of big data, as well as to gain the full benefit of the level of granularity and "realtimeliness" it offers.

---

65 Used for the creation of labour market indicators as it contains a labour force module.
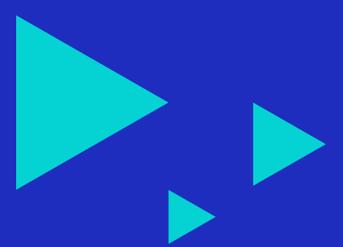
## ▶ Conclusion

First-generation applications of "big data" techniques to skills analysis have focused principally on analysis of large-scale collections of vacancies data, either scraped from multiple online sources or provided directly by jobs websites or employment agencies. Assembly, coding and cleaning of these data sets, and subsequent analysis and interpretation of the data, present significant challenges, but this type of application is progressing towards maturity. There are multiple competing vendors, covering a range of countries with different languages, providing services commercially to customers such as universities and employers. The technology is being adopted by multiple public LMI and skills anticipation systems to complement their existing methodologies and information products.

Information gained from applying big data techniques to large-scale vacancies data is a higher-resolution, bigger-data version of the analyses of print advertisements that were used for content analysis in some LMI systems in the past. It is not a direct substitute for the core data sources for LMI and skills anticipation, such as LFS. Even when the challenges presented by the data sources are resolved to the greatest extent feasible, the nature of the information provided is different from that provided by regular statistics.

For instance, a household survey conducted through a statistically well-founded sampling process provides a detailed snapshot of employment in an economy for the time at which the survey is conducted. It can also be used to derive information on change, both at aggregate level by comparing successive iterations of the survey, and at micro-level, through recording responses to questions on how the employment characteristics of individuals in each household survey have changed in the preceding month or year. The main proxy indicators that provide insights into workers' skills are those on occupation, sector of employment and qualifications. Age also provides a proxy for years of experience.

By contrast, vacancies data provides indicators on skills requirements associated with a combination of new jobs and churn in existing jobs. The indicators are imperfect in ways that are discussed in the contributions to this report, but at their best they can provide information on advertised vacancies sorted by occupations and qualifications, coded similarly to regular statistics, taking into account standard classifications such as ISCO (SOC), ISIC and ISCED. As standard national statistics rarely provide data on specific salient skills available and required, big data in OJV becomes a unique complementary source to fill out the up-to-date picture with information on specific job tasks, skills, qualifications and certifications valued, and experience demanded. Such analyses are already proving valuable to private users of OJV big data, and collections of indicators combining different information sources seems likely to become a regular feature of national LMI and skills anticipation systems as they integrate OJV into their processes and products.

In developing economies, especially in low-income countries, LMI on skills proxies such as occupation and level of education, let alone on specific skills and competencies, is often missing. Usually, such information is collected by qualitative and semi-qualitative methods, such as occupational research, expert committees or surveys among employers and workers. Such approaches may be rather costly, especially if the information collected is to provide the level of granularity needed for analysis capable of informing the design of education and training. In low-income countries, regular statistical sources of information, such as LFS, may not be conducted frequently or regularly enough. The use of big real-time data analytics may potentially resolve this problem.

However, access to the Internet and levels of digital literacy are important limitations in developing countries. Furthermore, given the large size of the informal economy in these countries, only a small portion of actual vacancies are covered by online job platforms, where it is very likely that only the formal sector is represented. These considerations indicate significant challenges facing the use of OJV data, particularly in developing countries. Therefore, once the focus shifts to developing countries, it is important to keep these limitations in mind and treat data analysis as valid in the specific context only.

Because of how job advertisements are written, the information on skills requirements available for extraction is better for technology jobs and high-skill white-collar jobs, thinner for mid-level jobs, and often marginal for low-level jobs. Vacancies not advertised online remain unobserved, typically meaning that

lower-skill vacancies and vacancies in less developed economies are under-represented, and that vacancies intended to be filled through internal organizational movements are largely absent. That is why, at least for the time being, big data is best used when analysed in relation to what it represents – specific occupations and sectors, mostly digitally intensive and more highly skilled.

Up to now, the main focus of skills-related big data activity has been on OJV. Circumstances have been favourable for this. The principal source of vacancies data, in the form of online job advertisements, is accessible without substantial restrictions or fees. Because the data do not relate to individuals, and are already published online, constraints associated with individual and commercial privacy have so far been minimal though they are likely to surface and require attention in the future. There is a commercial market in industrialized countries for vacancies analysis services – for example, for course planning and careers services at universities, for recruitment intelligence in major businesses and for analysis of financial instruments – so there is already a business model that does not rely on public policy applications.

However, a wide range of other types of big and "small" data resources on skills exists, going far beyond vacancies data, which could potentially be combined for analysis, subject to the satisfactory management of issues around access, privacy, individual identifiers and data protection. These may be summarized as follows:

▶ *Public administrative microdata:* Administrative data drawn from across ministries and other public organizations is increasingly used by national statistical offices to complement or replace survey data in the preparation of both statistics for publication and customized analyses in support of policy formation. Individual ministries and their agencies have access to their own administrative data, and may also negotiate access to administrative data held by other ministries for policy analysis. Most if not all countries have large holdings of public administrative data relevant to skills.

▶ *National statistical office microdata:* National statistical offices have huge microdata holdings, including data relevant to skills from statistical surveys, such as LFS, census of population and other household surveys; the US Occupational Requirements Survey; Eurostat's Adult Education Survey and Continuing Vocational Training Survey.

▶ *Large-scale skills survey and skills measurement microdata:* In many countries, public and private organizations beyond national statistical offices undertake substantial enterprise skills surveys, some of them as one-off events, others repeated periodically. Various others provide services such as large-scale analysis of CVs on behalf of recruiting companies, or large-scale measurement of skills for purposes of issuing certifications.

In this report, a number of studies and approaches have been presented, showing the multitude of contexts in which job vacancy big data can be used to gain a better understanding of the underlying dynamics of the labour market, in particular on the demand side. The studies range from more technical ones, focusing on the actual extraction and management of OJV, and how these can be rendered susceptible to any kind of analysis, to national and cross-country studies that combine big data analytics with other types of regular LMI. While the majority of experience so far comes from advanced economies, the studies of India, Myanmar, Latin America and the Caribbean prove that real-time big time data may open up new ways of better understanding the dynamics in emerging and developing economies and may compensate, to some extent at least, for the absence of standard statistics. The future will show to what extent big data analytics will be mainstreamed into national LMI systems, replacing or complementing other sources of information, and whether it will allow less developed countries that do not have established data systems to leapfrog into effective skills needs anticipation and matching.

# ▶ References

Alabdulkareem, A.; Frank, M.R.; Sun, L.; AlShebli, B.; Hidalgo, C.; Rahwan, I. 2018. "Unpacking the polarization of workplace skills", in *Science Advances*, Vol. 4, No. 7, eaao6030.

Aggarwal, C.C. ; Zhai, C. 2012. "A survey of text classification algorithms", in C.C. Aggarwal and C. Zhai (eds): *Mining text data* (New York, Springer), pp. 163–222. Available at: http://link.springer.com/chapter/10.1007/978-1-4614-3223-4_6.

Auzinger, M.; Broek, S.; McCoshan, A.; Hudepohl, M.; Luomi-Messerer, K.; Plaimauer, C. Forthcoming. *Exploring gathering and analysing national qualifications data: Comparing VET qualifications*, work assignment 2 draft final report, 31 Oct. 2019. Available at: https://www.cedefop.europa.eu/en/events-and-projects/events/workshop-comparing-vet-qualifications.

Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. 2003. "A neural probabilistic language model", in *Journal of Machine Learning Research*, Vol. 3, Feb., pp. 1137–1155.

Berg, A.M.; Branka, J.; Kismihók, G. 2018. "Combining learning analytics with job market intelligence to support learning at the workplace", in D. Ifenthaler (ed.): *Digital workplace learning* (Cham, Springer), pp. 129–148. Available at: https://doi.org/10.1007/978-3-319-46215-8_8.

Berkers, H.; Mol, S.T.; Kobayashi, V.; Kismihók, G.; Hartog, den D. 2019. "Big (data) insights into what employees do: A comparison between task inventory and text mining job analysis methods", in H. A. Berkers, *What do you do and who do you think you are? Activities speak louder than words*, PhD thesis, University of Amsterdam, pp. 12–57. Available at: https://pure.uva.nl/ws/files/31377407/Chapter_2.pdf.

Bongers, F.; Jager, C-J.; te Velde, R. 2015. *Big data in onderwijs en wetenschap: Inventarisatie en essays* (Utrecht, Netherlands Ministry of Education).

Boselli, R., Cesarini, M., Mercorio, F. and Mezzanzanica, M. 2015, "A model-based evaluation of Data quality activities in KDD", in *Information Processing and Management*, Vol. 51, No. 2, pp. 144–166.

—; Cesarini, M.; Mercorio, F. Mezzanzanica, M. 2017. *Using machine learning for labour market intelligence*, paper presented at "Machine learning and knowledge discovery in databases", European conference, ECML PKDD 2017, Skopje, Macedonia, 18–22 Sep., *Proceedings*, part III, pp. 330–342.

—; Cesarini, M.; Marrara, S.; Mercorio, F.; Mezzanzanica, M.; Pasi, G.; Viviani, M. 2018. "Wolmis: A labour market intelligence system for classifying web job vacancies", in *Journal of Intelligent Information Systems*, Vol. 51, No. 3, pp. 477–502.

Cappelli, P.; Tambe, P.; Yakubovich, V. 2019. "Artificial intelligence in human resources management: Challenges and a path forward", in *SSRN Electronic Journal*, publ. online 8 Apr. Available at https://ssrn.com/abstract=3263878 or http://dx.doi.org/10.2139/ssrn.3263878.

Cedefop. 2014. *Real-time labour market information on skill requirements: Feasibility study and working prototype* (Thessaloniki).

—. 2016. *Real-time labour market information on skill requirements: Setting up the EU system for online vacancy analysis* (Thessaloniki).

—. 2019a. *The online job vacancy market in the EU: Driving forces and emerging trends* (Thessaloniki). Available at: https://www.cedefop.europa.eu/files/5572_en.pdf.

—. 2019b. *Online job vacancies and skills analysis: A Cedefop pan-European approach* (Thessaloniki). Available at: https://www.cedefop.europa.eu/files/4172_en.pdf.

Colombo, E.; Mercorio, F.; Mezzanzanica, M. 2019. "AI meets labour market: Exploring the link between automation and skills", in *Information Economics and Policy*, Vol. 47, No. C, pp. 27–37.

Curry, B. 2014. "Job-vacancy rate plunges as Tories drop Kijiji data", *Toronto Globe and Mail*, 5 May. Available at: https://www.theglobeandmail.com/report-on-business/economy/jobs/ottawa-adjust-labour-data-raising-questions-about-national-skills-shortage/article18457198/.

Deming, D.; Khan. L.B. 2017. *Skill requirements across firms and labor markets: Evidence from job postings for professionals* (Cambridge, MA: National Bureau of Economic Research).

de Pedraza, P.; Visintin, S.; Tijdens, K.; Kismihók, G. 2019. "Survey vs scraped data: Comparing time series properties of web and survey vacancy data", in *IZA Journal of Labor Economics*, Vol. 8, No. 1. Available at: https://doi.org/10.2478/izajole-2019-0004.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. 1996. "The KDD process for extracting useful knowledge from volumes of data", in *Communications of the ACM*, Vol. 39, No. 11, pp. 27–34.

Frey, C.B.; Osborne, M.A. 2017. "The future of employment: How susceptible are jobs to computerisation?", in *Technological Forecasting and Social Change*, Vol. 114, Supplement C, pp. 254–280.

International Labour Office (ILO). 2012. International Standard Classification of Occupations: Structure, group definitions and correspondence tables (Geneva). Available at: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_172572.pdf [28 July 2020].

—. 2018. *World Employment and Social Outlook: Greening with Jobs* (Geneva).

—. 2019. *Skills for a greener future: A global view based on 32 country studies* (Geneva).

—. Forthcoming 2020. *Global Employment Policy Review: Employment policies for inclusive structural transformation* (Geneva).

Karsten, S. 2016. *De hoofdstroom in de Nederlandse onderwijsdelt*a (Antwerp, Apeldoorn).

Kobayashi, V.; Mol, S. .; Kismihok, G.; Hesterberg, M. 2016. "Automatic extraction of nursing tasks from online job vacancies", in M. Fathi, M. Khobreh and F. Ansari (eds): *Professional education and training through knowledge, technology and innovation* (Siegen, Universitätsverlag Siegen), pp. 51–56. Available at: http://www.pro-nursing.eu/web/resources/downloads/book/Pro-Nursing_Book.pdf.

—; Mol, S.T.; Berkers, H A.; Kismihók, G.; Den Hartog, D.N. 2017a. "Text mining in organizational research", in *Organizational Research Methods*, publ. online 10 Aug., https://doi.org/10.1177/1094428117722619.

—; Mol, S.T.; Berkers, H.A.; Kismihók, G.; Den Hartog, D.N. 2017b. "Text classification for organizational researchers: A tutorial", in *Organizational Research Methods*, publ. online 12 July, https://doi.org/10.1177/1094428117719322.

Mezzanzanica, M.; Mercorio, F. 2019a. *Big data for labour market intelligence: An introductory guide* (Turin: European Training Foundation). Available at: https://www.etf.europa.eu/sites/default/files/2019-06/Big%20data%20for%20LMI.pdf.

—; —. 2019b. *Feasibility study for Tunisia and Morocco to identify, validate and rank web job vacancy sources: Practical guidance*, unpubl. paper (Turin, European Training Foundation).

—; —. 2019c. "Big data enables labour market intelligence", in S. Sakr and A. Zomaya (eds): *Encyclopedia of Big Data Technologies* (Cham, Springer).

Mikolov, T.; Yih, W.; Zweig, G. 2013. "Linguistic regularities in continuous space word representations", in *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, Atlanta, GA, June (Stroudsberg, PA, Association for Computational Linguistics).

Nania, J.; Bonella, H.; Restuccia, D.; Taska, B. 2019. *No longer optional: Employer demand for digital skills* (London, Burning Glass Technologies for Department for Digital, Culture, Media and Sport).

Organisation for Economic Co-operation and Development (OECD). 2017. *Getting skills right: Skills for jobs indicators* (Paris). Available at: https://www.oecd-ilibrary.org/employment/getting-skills-right-skills-for-jobs-indicators_9789264277878-en [28 July 2020].

Plaimauer, C. 2015. *Validation of OSPs [occupational skills profiles] using semantic technologies*, final report for Cedefop, FWC No. AO/RPA/AZU-VKVET/skill-forecast/003/12, led by Warwick Institute for Employment Research.

—. 2016. *Big Data als Informationsquelle für regionales Arbeitsmarkt-Monitoring: Online-Stellenanzeigen analysieren mittels "Jobfeed"*, AMS info No. 339 (Vienna, AMS Forschungsnetzwerk). Available at: http://www.forschungsnetzwerk.at/downloadpub/AMS_info_339-1.pdf.

—. 2018a. *Die Darstellung der Arbeitskräfte-Nachfrage im AMS-Qualifikations-Barometer*, AMS info Nos 410/411 (Vienna, AMS Forschungsnetzwerk). Available at: http://www.forschungsnetzwerk.at/downloadpub/AMS_info_410-411.pdf.

—. 2018b. "Monitoring changing needs in skills, competences and qualifications in Austria", in C. Larsen, S. Rand, A. Schmid and A. Dean (eds): *Developing skills in a changing world of work: Concepts, measurement and data applied in regional and local labour market monitoring across Europe* (Augsburg, Rainer Hampp).

Redman, T.C. 1998, "The impact of poor data quality on the typical enterprise", in *Communications of the ACM*, Vol. 41, No. 2, pp. 79–8

Rodenberg, J.H.A.M. 2002. *Business and competitive intelligence* (Alphen aan den Rijn, Kluwer).

Solka, J.L. 2008. "Text data mining: Theory and methods", in *Statistics Surveys*, Vol. 2, No. 0, pp. 94–112. Available at: http://doi.org/10.1214/07-SS016.

UK Commission for Employment and Skills. 2015. *The importance of LMI* (London). Available at: https://www.gov.uk/government/publications/the-importance-of-labour-market-intelligence.

UK Office of National Statistics (ONS). 2015. *Annual survey of hours and earnings* (London).

UNESCO. 2019. TVET system review: Myanmar (Paris). Available at: https://unesdoc.unesco.org/ark:/48223/pf0000371392.

Wang, R.Y. and Strong, D.M. 1996, "Beyond Accuracy: What Data Quality Means to Data Consumers", in *Journal of Management Information Systems*, Vol. 12, No. 4, pp. 5–33.

Winterbotham, M.; Vivian, D.; Kik, G.; Hewitt, J.H.; Tweddle, M.; Downing, C.; Thomson, D.; Morrice, N.; Stroud, S. 2018. *Employer skills survey 2017: Research report* (London: UK Department for Education. Available at: https://www.gov.uk/government/publications/employer-skills-survey-2017-uk-report.

Winzenried, S. 2018. *Even more ado about nothing ... or why the hype about big data and AI is often more about self-marketing than facts and real progress*, 22 Aug. (Zurich, Janzz Technology). Available at: https://janzz.technology/even-ado-nothing-hype-big-data-ai-often-self-marketing-facts-real-progress/.

—; Zanoli, J. 2016. *Lost in big data? The misguided idea ruling the data universe*, 3 May. Available at: https://janzz.technology/lost-in-big-data/.

Wissner-Gross, A. 2016. "Datasets over algorithms", in Edge 2016: *What do you consider the most interesting recent (scientific) news? What makes it important?* (New York, Brockman). Available at: https://www.edge.org/response-detail/26587.

World Economic Forum (WEF). 2018a. *The 10 skills you need to thrive in the Fourth Industrial Revolution* (Cologny). Available at: https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/

—. 2018b. *The future of jobs report* (Cologny). Available at: https://www.weforum.org/reports/the-future-of-jobs-report-2018.

—; Boston Consulting Group (BCG). 2019. *Towards a reskilling revolution: Industry-led action for the future work* (Geneva). Available at: http://www3.weforum. org/docs/WEF_Towards_a_Reskilling_Revolution.pdf.

Zhu, T.J.; Fritzler, A.; Orlowski, J.A.K. 2018. *World Bank Group-LinkedIn data insights: Jobs, skills and migration trends methodology and validation results*, Vol. 2: Executive summary (English) (Washington, DC). Available at: http://documents.worldbank.org/curated/en/561481542144877029/Executive-Summary.

Ziegler, P.; Auzinger, A. 2012. AMS *Großbetriebsmonitoring 2011: Bericht und Auswertungen zu nachgefragten und dynamischen Berufen und Qualifikationen* (Vienna, AMS Forschungsnetzwerk). Available at: http://www.forschungsnetzwerk.at/downloadpub/2012_ams_Bericht_Gro%C3%9Fbetriebs_Monitoring_2011.pdf.