CEDEFOP | European Centre for the Development of Vocational Training

eurostat

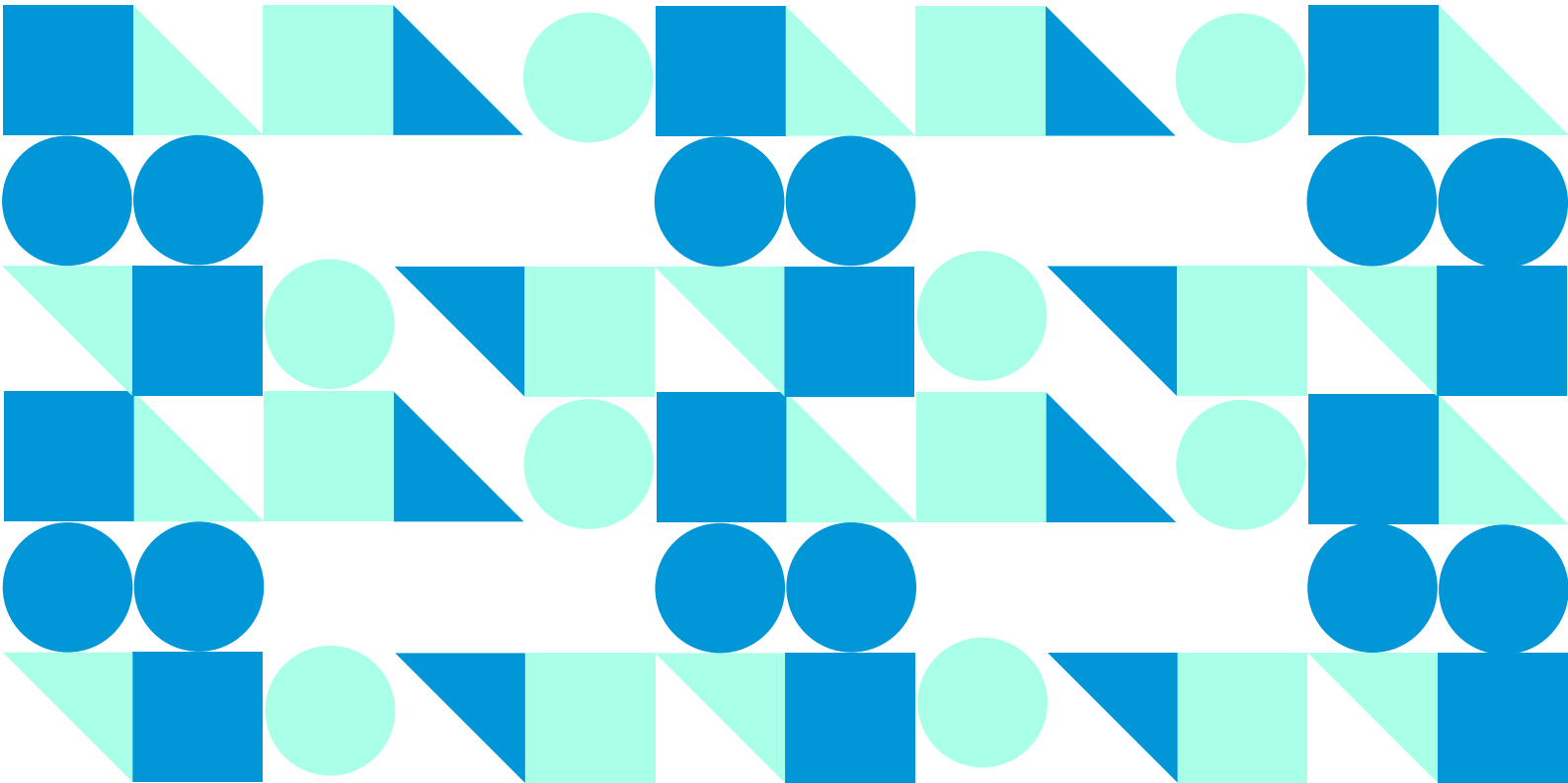
EN

1831-5860

Research paper

Delivering evidence from online job advertisements

Tapping into 10 years of experience



CEDEFOP

Delivering evidence from online job advertisements

Tapping into 10 years of experience

Luxembourg: Publications Office of the European Union, 2025

Please cite this publication as:
Cedefop & Eurostat (2025). *Delivering evidence from online job advertisements: tapping into 10 years of experience*. Cedefop research paper. Publications Office of the European Union. http://data.europa.eu/doi/10.2801/5070484

PDF ISBN 978-92-896-3841-8
ISSN 1831-5860
doi: 10.2801/5070484
TI-01-25-049-EN-N

The **European Centre for the Development of Vocational Training** (Cedefop) is the European Union's reference centre for vocational education and training, skills and qualifications. We provide information, research, analyses and evidence on vocational education and training, skills and qualifications for policymaking in the EU Member States. Cedefop was originally established in 1975 by Council Regulation (EEC) No 337/75. This decision was repealed in 2019 by Regulation (EU) 2019/128 establishing Cedefop as a European Union agency with a renewed mandate.

Europe 123, Thessaloniki (Pylaia), Greece
Postal: Cedefop service post, 57001 Thermi, Greece
Tel. +30 2310490111, fax +30 2310490020
Email: info@cedefop.europa.eu
www.cedefop.europa.eu

Eurostat is the statistical office of the European Union (EU). Its mission is to provide high-quality statistics and data on Europe, enabling comparisons between countries and regions. Operating independently within the European Commission, Eurostat develops harmonised definitions, classifications, and methodologies to produce European official statistics. In partnership with National Statistical Institutes and other national authorities, Eurostat ensures the collection, analysis, and dissemination of accurate and consistent statistical data across EU Member States.

Foreword

Granular and timely information on skills is key for designing and implementing effective and inclusive education and training policies. European and national labour markets also rely heavily on such 'skills intelligence' to enable and shape fair and effective green and digital twin transitions and to mitigate the quantitative shortages of human capital caused by demographic change.

Achieving consistent, comparable and reliable data from conventional skills anticipation approaches, such as skills forecasts or surveys, comes at a price. Typically, information is available only at higher aggregation levels and following long time lags, and skills are not directly measured but proxied for by other concepts, such as occupations or qualifications. Over the past decade, the European Centre for the Development of Vocational Training (Cedefop) and Eurostat, the statistical office of the European Union, have teamed up to champion and explore new data sources that can help to overcome these issues. As online platforms have become the primary way of looking for talent, online job advertisements (OJAs) give an important insight into the skills employers want. While the internet has provided a rich source of information for a long time, thanks to recent advances in cloud computing and machine learning we can now fully leverage the richness of these data to develop policy-relevant intelligence on jobs, skills and labour market trends.

Cedefop launched its first OJA-powered skills intelligence as a pilot for seven EU Member States in 2019, following several years of building a system from the ground up. Since then, a lot of investment, work and effort has gone into setting up and quality-assuring a pan-European system to collect and analyse OJAs. It was obvious from the start that OJAs contain a wealth of information that can be leveraged to map labour market and skills trends. At the same time, analysing data from sources where the primary aim is recruiting staff – not research and analysis – brings a host of theoretical and practical challenges.

Although interpreting skills intelligence based on OJAs can be complex and challenging, Cedefop decided to make indicators and dashboards publicly available from the start. Over time, Cedefop's Skills-OVATE data visualisation web tool developed a reputation for offering novel data on skills trends. By 2023, the tool had been expanded to include additional countries, and in 2024 twin transition dashboards were introduced to give users more detailed insights into the skills impact of the digital and green transitions.

In the past few years, Cedefop and Eurostat have worked in close partnership on OJA data. Eurostat has promoted the work of the European Statistical System

on the exploration of OJA data for statistical purposes since 2016, with a series of projects dedicated to the exploration of big data sources. The collaboration between Cedefop, Eurostat and the European Statistical System started with a series of joint workshops, knowledge-sharing efforts and the decision by Eurostat to use Cedefop's OJA data production system while developing its own system for the production of official statistics. The continuing partnership between Eurostat and Cedefop culminated in the creation of Eurostat's Web Intelligence Hub (WIH). This arrangement ensures the continuous development of high-quality web-based skills intelligence. Eurostat provides the infrastructure, produces the data, oversees and improves data quality and develops the statistical use of OJA data, with the first experimental statistics published in 2023 on the labour market demand for specialists in information and communications technology. Cedefop uses the data to provide evidence-based insights into skills dynamics in the labour market and the impact of megatrends on jobs and skills in sectors, countries and regions.

Ten years after the first feasibility study on using OJAs for EU skills intelligence, it is time to take stock. In this report, we map the progress we have made, we report on the methodology and we explain how the OJA data production system gathers and produces OJA data. We also provide insights into data quality management and data extraction methodologies and techniques, and we review and showcase how OJA data can be used in thematic analysis.

With this publication, we also want to signal our commitment to further developing big-data-based skills intelligence for Europe. Together, Cedefop and Eurostat will continue their joint journey of mutual learning and community building in support of better skills data. In the coming years, we will be working in close partnership to deliver user-centred insights into skills in jobs, occupations and sectors. This will support the efforts of countries and regions to strengthen the human and social capital in their societies and the competitiveness of their industries and economies. It will also contribute to shaping the skills revolution the EU needs to become greener, more digital and more resilient in a context of ageing populations, rapid AI and tech development and increasing uncertainty and volatility.

Jürgen Siebel
Cedefop Executive Director

Sophie Limpach
Director of Resources of Eurostat

Acknowledgements

This publication is the result of cooperation between Cedefop's Department for VET and Skills and Eurostat's Unit for Methodology and Innovation in Official Statistics as part of the project based on the framework contract 2020-FWC7/AO/DSL/VKVET-JBRAN/WIH-OJA/002/20, 'Towards the European Web Intelligence Hub – European system for collection and analysis of online job advertisement data (WIH-OJA)'.

Vladimir Kvetan, Joanna Napierala, Jiri Branka (all Cedefop), Raquel Tello de Faria Paulino and Anca-Maria Nagy (both Eurostat consultants from Sogeti) drafted this publication under the supervision of the Head of the Department for VET and Skills, Antonio Ranieri. Adam Tsakalidis (Cedefop), Giovanni Russo (Cedefop) and Fernando Reis (Eurostat) reviewed it.

Cedefop would like to acknowledge that the research consortium led by CRISP, the Interuniversity Research Centre for Public Services of the University of Milano Bicocca, under the leadership of Mario Mezzanzanica and Emilio Colombo, contributed to meeting the project's objectives.

Contents

Tables and figures

Tables

Figures

Executive summary

Introduction

Cedefop has been developing skills intelligence to support labour market analysis for two decades, with the aim of informing education and training policies in the EU. Most methods are based on 'conventional' approaches, largely based on official statistics, dedicated surveys and other data collections suitable for developing robust skills analysis and anticipation tools such as forecasts, scenario analysis and composite indexes. However, these approaches often lack the desired level of granularity and timeliness and, most of all, do not offer direct measurements of skills concepts. Primarily relying on proxies for skills, such as occupations or formal qualification levels, conventional data and approaches may lead to gaps in our understanding of individual skills needs in workplaces.

To bridge this gap, Cedefop has pursued its aim of delivering more timely and granular information on employers' skills requirements using new methods and data sources. Recognising the potential of online job advertisements (OJAs) as a real-time data source, Cedefop launched a feasibility study in 2015-16 to assess the viability of collecting and analysing OJA data. The success of this study, highlighted during the 2017 European Big Data Hackathon, encouraged further investment in a fully operational data production system (DPS) for OJA, which had become fully functional by 2020. The launch of Skills-OVATE facilitated data navigation, and by 2023 the system had expanded to include data for additional countries (Iceland, Liechtenstein, Norway and Switzerland) and gradually incorporated new data sources and variables.

In parallel, Eurostat, the statistical office of the European Union, initiated the ESSnet big data project in 2016 to integrate big data into official statistics, including OJA data. A collaboration between Cedefop and Eurostat led to joint workshops, knowledge sharing and, eventually, the recommendation to use Cedefop's OJA data system for official statistics. This partnership resulted in the integration of the Cedefop OJA DPS into the Web Intelligence Hub (WIH), ensuring continuous development and high-quality skills intelligence.

This publication maps a decade of progress in using OJA data, transitioning from basic data collection to detailed analyses of occupations and skills. It provides a structured overview of OJA-based labour market intelligence. Chapter 2 lays the foundation with a landscaping exercise, providing a crucial framework for understanding the OJA market and interpreting the results of analyses effectively. Chapter 3 describes the inside of the DPS, detailing the process of gathering and

analysing OJA data while outlining innovative methods for measuring and improving data quality. A more detailed description of extracting skills and occupations is given in Chapter 4, which also highlights system advances.

Understanding the value and tackling the shortcomings of the OJA data and methods used to classify occupations and skills drove us to develop a more detailed analysis. In particular, Chapter 5 sharpens the classification of professions and digital skills, while Chapter 6 explores the pivotal role of OJA data in monitoring the green transition. Examining the value of OJA data beyond providing information on skills and occupations, Chapter 7 expands the scope of requirements beyond skills and occupations to education fields and qualifications, enriching our understanding of their relevance to the labour market. Finally, Chapter 8 brings everything together, focusing on the ultimate goal – delivering cutting-edge skills intelligence derived from web data, particularly OJA data. This chapter underscores how skills intelligence can support labour market analysis and evidence-based policy.

Understanding the landscape of online job advertisements

OJAs are a rich source of labour market insights, despite being primarily designed to attract candidates rather than to generate statistical data. However, to fully understand the context, content and coverage of OJA data, conducting a 'landscaping exercise' before actually starting to collect data is considered fundamental. The exercise aims to identify and characterise the existing sources of OJAs in each country with features including business model, key source characteristics (e.g. content, frequency of updating) and even country regulatory framework or linguistic peculiarities. The main qualitative work was done by a group of individual country experts (ICEs) with detailed knowledge of the local online labour market. The analytical hierarchy process was used to organise and analyse complex information through a structured hierarchy of criteria.

Gathering and analysing online job advertisements through the data production system

The collection and analysis of the OJAs are organised through a multilingual modular DPS, thoroughly described in Cedefop (2019a). The structure of the DPS allows it to process OJAs in the original language of the document. The web sources of OJAs identified through the landscaping exercise (described in

Chapter 2) enter the ‘data ingestion’ phase. In this phase, the system downloads the content of individual OJAs. The downloaded documents are cleaned of irrelevant content during the next phase – ‘preprocessing’. Subsequently, the ‘information extraction’ phase leads to the production of data categorised by individual variables: place of work, occupation, economic activity, etc.

The ‘data validation’ phase is essential for identifying errors, refining classification models and enhancing machine learning processes to create high-quality output from OJAs. While validation does not directly measure overall data quality, it is crucial in improving dictionaries, ontologies and classifiers. Validation is conducted iteratively. The process employs validation rules to check compliance with official classifications, detect anomalies and enforce data integrity. Adapting to evolving requirements and technological advances, the validation framework remains dynamic, reinforcing data reliability and precision.

Extracting skills and occupations

The ultimate goal of the processes organised within the DPS is to extract information from OJAs. Among the many variables we are able to extract (e.g. place of work, economic activity, type of contract), the occupation and skills variables are vital for further combination with other ‘conventional’ data sources (such as surveys and/or forecasts). The DPS employs a blend of machine learning and ontology-based methods to extract information to ensure that the results are robust and accurate. While machine learning offers better and more flexible results, the ontology-based approach helps to address the DPS’s multilingualism and provides more stable and comparable data extractions across time.

The occupations are extracted and classified based on the 2008 International Standard Classification of Occupations (ISCO-08), assigning labels to OJAs at the unit group level (ISCO-08 four-digit) and offering highly granular information on occupations. Large language models (LLM) could bring even more granular and more precise results in the classification of job titles. However, there is a clear issue with resources in terms of volume and the predictability of costs. Initial testing has confirmed that using a ‘direct approach’ (prompting an LLM to assign an occupation code to an OJA job title) within the current WIH DPS takes about three seconds (for more technical details, see Annex 1). Considering the scale of the task, the total time required would be impractically long and difficult to predict.

Therefore, an alternative strategy has been developed based on semantic representations of OJAs and occupations. This approach involves leveraging an LLM to calculate embeddings for each OJA and each occupation only once. The embeddings for the OJAs and the occupations were pre-computed and stored. The

exercise yielded positive results for occupation classification, speed and cost. However, from the perspective of the WIH DPS, it is essential to assess whether this LLM-based method is more accurate than the currently prevalent ontology or fuzzy matching. At the ISCO four-digit level, the improvement is notable. The LLM-based classification system is also an improvement on broader categorisation models by accurately classifying jobs that would have previously fallen into a catch-all 'not elsewhere classified' category. Instead of being lumped together under a generic classification, many of these jobs are recognised for their specialised functions, one of the exercise's most desired outcomes. The current pilot, however, needs to be put into production and tested for consistency over time before the DPS will start officially publishing data on detailed European Skills, Competences and Occupations (ESCO) classification system occupations.

Skills classification relies on ontology matching, utilising ESCO as an overarching taxonomy. Extracting skills from OJAs, however, presents considerable complexities. Unlike job titles, which tend to be explicit in OJA titles, skills needs are usually expressed in the unstructured sections of job advertisements. They are articulated in many ways and greatly influenced by linguistic nuances. Despite the breadth of ESCO's alternative labels, these may not always align with the specific language employers use in job postings. Often, certain skills are not overtly requested, as they are presumed to be inherent to the job role. Employers may also rely on the stated qualification requirements as proxies for various skills, eschewing the need to list each skill individually. The assessment of skills captured across countries, languages and occupations is one of the regular activities of the quality monitoring framework.

Alongside various other regular activities, two more detailed activities were devised to understand the issues in extracting skills. The first focused on augmentation following the manual translation of skills terms that most frequently occur in English to all languages of the DPS. The second utilised machine learning and a natural language processing method to extract frequently occurring terms that are close to skills already extracted. The ICEs then assessed the terms to determine whether or not they represented a skill. The LLM was used to assist the ICEs in decision-making. Both activities contributed to improving the ESCO-based ontology on skills.

Using online job advertisements to understand the digital transition

The EU prioritises understanding the digital transition's impact on the labour market and the skills needed. However, the rapid evolution of digital technology

makes it challenging to keep classification systems like ESCO up to date. Although ESCO updates, like the recent version 1.2, include skills for tools like ChatGPT, the process of including new skills is often slow and costly, as it requires consultation and the involvement of national experts. This delay highlights the need for more timely updates, particularly as the demands for digital skills increase.

Two platforms, Stack Overflow and GitHub, were analysed as potential sources for identifying and classifying new digital skills. Stack Overflow's community-driven tags system labels technology topics, and GitHub's repository tags provide information on technology usage and trends. The combined data from these platforms initially included around 65 000 terms, later refined to 40 000 through deduplication.

LLMs were used to measure the similarity between new digital terms and ESCO's digital skills subset to align these tags with existing ESCO classifications. The models ranked similarity matches using a quartile-based method, reducing the list to terms that showed the most substantial alignment with the ESCO classification. Further filtering identified which terms were directly relevant to digital skills.

To refine term selection further, the analysis cross-checked these terms with UK OJAs to assess the demand for them on the job market. Terms with little relevance or generic digital terms were removed, leaving a list that leaned heavily towards software tools, networks and programming languages. A panel of domain experts evaluated 1 733 terms for inclusion in the ESCO classification. Terms were accepted if they represented concrete skills or knowledge relevant to digital professions. About 23% of terms were considered valid, with a substantial portion related to software tools and computer networks.

The process demonstrated that integrating OJA data and LLMs can streamline and update digital skills classifications. Using community-driven platforms such as Stack Overflow and GitHub, alongside expert input, helped identify a relevant set of emerging digital skills ready to be proposed for ESCO updates.

Using online job advertisements to understand the green transition

The European Green Deal has set ambitious goals to make the EU climate neutral by 2050, which will require a significant transformation in the workforce through the development of 'green skills' tailored to new, environmentally sustainable industries. Chapter 6 explores how data from OJAs can help identify these

emerging skills and occupations, facilitating a smoother transition by informing education and training policies.

A bottom-up, data-driven approach was designed to analyse the skills requirements associated with green jobs through advanced machine learning and natural language processing techniques. Initially, the project aimed to create a unique classification of green skills based solely on OJA data, as a formal classification was not publicly available. However, the release of the ESCO green skills and knowledge concepts classification allowed researchers to test their extracted terms for alignment with ESCO's framework. Through these tests, they identified and proposed additional green skill terms to refine the classification further.

There is no universal standard for defining green jobs or skills. Building on earlier occupational network frameworks like the O*NET green economy programme, this study focused on green skills tied to specific processes, products, industries and occupations, using skills to determine how green the occupation is. A machine learning model, trained on 140 key terms drawn from sources like the Classification of Environmental Protection Activities, International Renewable Energy Agency, O*NET and LinkedIn, was applied to millions of UK-based OJAs. The model leveraged distributional semantics and word embedding methods to identify new green skills with high accuracy. For example, the terms 'ecological' and 'sustainable' yielded over 37 new green skill proposals when matched with OJA text.

Experts from various countries translated and validated the skills terms, ensuring that the green skills identified aligned with their occupational contexts. This process allowed the list to grow from 140 to 182 green skills terms, expanding its reach to cover multiple European languages by 2023.

Green occupations sometimes lack explicit green skills terms in job descriptions, especially when the green element is implied in the job title (e.g. environmental engineer). Additionally, non-skill-related green terms, such as those describing company missions, can introduce potential noise in the data. To mitigate this, specific filters were used to limit the analysis to skill-related text only.

Extracting information about the field of study

Despite the recent emphasis on skills-based hiring, qualifications remain essential in hiring processes. Qualifications help employers screen candidates, mitigate hiring risks and ensure compliance with regulatory requirements. Cedefop has analysed how employers specify required fields of study to better understand how

qualifications are expressed in job advertisements. Cedefop has explored two primary methods for extracting field-of-study information from OJAs.

The first, an ontology-based extraction method, relies on the International Standard Classification of Education – Fields of education and training (ISCED-F) framework. While effective, its limitations include inconsistent adoption across languages and variations in the number of available terms across different countries. The second, a data-driven approach, identifies 'anchoring terms' near field of study references in job postings. Terms such as 'degree' or 'qualification' help extract study field information. Validation by experts helped refine the list, improving extraction accuracy. However, the results varied across countries due to sample size limitations.

A comparative analysis of the methods highlights several key findings. The ontology-based approach provides a structured and standardised way of classifying qualifications, ensuring alignment with international frameworks. However, its effectiveness is hindered by language inconsistencies and limited adoption of ISCED-F classifications across countries. The data-driven approach allows a more flexible and dynamic extraction process, adapting to how employers phrase qualification requirements, but it relies on expert validation to refine the accuracy of the terms extracted.

The validation process conducted by Cedefop's ReferNet experts played a crucial role in improving the reliability of the data-driven method. Despite this, variations in sample sizes across languages affected the robustness of the results in certain countries. In countries with larger datasets, such as Spain, Cyprus and Portugal the results were more accurate using the data-driven approach. In comparison, smaller datasets in nations like Denmark, Slovenia and Finland yielded less accurate results. Integrating national language terms with anchoring words in the extraction process demonstrated that combining multiple approaches enhances accuracy, but gaps remain where validation data are insufficient.

Developing skills intelligence from online job advertisements

Cedefop defines skills intelligence as 'the outcome of an expert-driven process of identifying, analysing, synthesising and presenting quantitative and/or qualitative skills and labour market information. These may be drawn from multiple sources and adjusted to the needs of different users.' As the content of OJAs captures the latest trends in occupations in demand and skills required and reflects employers' preferences, they offer a valuable source of information that can be used for building skills intelligence.

Converting OJA data into skills intelligence is a complex task involving advances in big data analytics, machine learning and natural language processing methods. Despite that, the information extracted from the body of OJAs has already been used to contribute to various types of labour market analysis. The data have primarily contributed to analysing skills needs in occupations, shedding light on labour and skills shortages within specific sectors and occupations, or short-term forecasting of skills trends. In recent years, Cedefop has used OJA data to better understand various dimensions of the impact of the twin transitions, as reflected in the changing demand for skills and occupations.

To understand which occupations and skills are in demand, Cedefop has utilised the granularity of information contained in OJAs. However, as the occupation structure (ISCO) is usually updated once every 20 years (the last one was in 2008), it may not capture the changing trends in emerging occupations. The key jobs for the twin transitions are often assigned to a broad group of 'occupations not elsewhere classified'. Classifying OJAs at the level of individual ESCO occupations (one level of disaggregation lower than ISCO) may allow better insights into the occupations or skills in demand. We believe that such granular information may help employers or training providers to tailor their programmes better.

The rapid pace of technological change is reshaping industries and transforming the nature of work. This evolution creates new job roles while eliminating some traditional ones and requires workers to adapt and acquire new skills. This translates into the need for skills intelligence to address questions related to emerging occupations and their skills requirements. Cedefop used OJAs to identify the skills required for niche sectors like blockchain technology or hydrogen energy.

At the same time, to address the gap in monitoring the impact of the green transition on the labour market and better understand how each occupation is affected by the green transition by using OJAs, we constructed two indicators: green pervasiveness and greenness. Green pervasiveness measures the prevalence of green skills in the OJAs. It is calculated as the ratio of all OJAs with at least one green skill requested by employers to the total number of OJAs in the category analysed (e.g. at the occupational, sectoral and country levels). The greenness indicator compares the number of green skills requested by employers with the overall number of skills found in the advertisements in the category analysed.

Similarly to greenness and green pervasiveness, we constructed indicators that will help us assess the impact of the digital transition on the skills required by employers. For example, digitalness measures how essential digital skills are for

a single occupation, and is defined as the ratio of the number of digital skills required to the total number of skills required, while digital pervasiveness measures the percentage of OJAs that demand at least one digital skill. Yet, in the case of digital skills, one might also be interested in understanding how digitalisation translates into the demand for various levels of skills.

The content of OJAs may also serve as a valuable tool for monitoring skills changes at the sectoral level, for example the speed and range of adoption of circular economy principles by employers across Europe. The analysis of skills mentioned in OJAs can also be used to detect the impact of introducing green or other policies on selected occupations and their required skills. Developing a measure of the change in the skill sets of an occupation is not easy. The skill set can change because some skills become more or less critical and because new skills become essential for an occupation. For example, as businesses and organisations are motivated by various policies to integrate sustainability principles into their operations, regardless of their primary focus, we may observe some changes in the skills requirements for these roles.

Conclusions

Over the past decade, Cedefop and Eurostat have built a cutting-edge, multilingual DPS for analysing OJAs, revolutionising labour market intelligence. By combining web scraping, machine learning and ontology-based classification, DPS filters and refines vast amounts of job posting data, allowing the extraction of detailed insights into skills demand, qualifications and emerging labour market trends. The study also demonstrates how OJA data may enhance skills frameworks (e.g. ESCO), using the example of digital skills, while addressing challenges in capturing green skills through innovative text-filtering techniques. However, cross-country comparability remains challenging due to linguistic and cultural differences, underscoring the need for a hybrid approach that blends automation with expert validation.

More than just a data pipeline, the OJA DPS is a game changer for policymakers, educators and employers, enabling smarter workforce planning and skills development. In an era of rapid technological change, the ability to decode job market signals from data is not just an advantage – it is a necessity.

Chapter 1.

Introduction

The European Centre for the Development of Vocational Training (Cedefop) has been developing skills intelligence with a view to supporting education and training policies in the EU for two decades ([1]). During this time, Cedefop has succeeded in developing and regularly updating pan-European skills forecasts and the European Skills Index (ESI) and has worked on different skills surveys, including the European skills and jobs survey. These activities provide essential input to pursuing a better match between skills supply and demand in the EU. Although Cedefop's work refers to skills intelligence, skills analysis has primarily relied on various proxies for skills rather than actual skills or knowledge concepts.

For example, skills forecasts use occupations and broad levels of formal qualifications ([2]). The European Skills Index, Cedefop's composite indicator, focuses on measuring the performance of EU skills systems. The index combines various official indicators in three pillars: skills development, skills activation and skills matching. Each indicator measures a different aspect of a skills system, but it refers to skills only indirectly. The skills surveys focus either on general measures of subjective skills engagement or on specific broadly defined types of skills, such as lifting loads, caring for others or using specialised software. Thus, an understanding of the individual skills needs of employers has been missing.

To understand individual skills needs and gaps within workplaces, Cedefop examined the potential of using employers' surveys. The goal was to create a reliable tool for assessing the skills, competences, occupations and qualifications required in European public and private enterprises, aiding broader analyses of skills needs. While various methods and sources can provide such insights, employers' surveys offer a qualitative and quantitative perspective. Although the pilot project was considered a success, the move to a full-scale survey was difficult due to the inability to provide sufficient detail on skills and occupations using the financial resources available.

The need to deliver timely and granular information on the skills required by employers has prompted Cedefop to seek new methods and data sources. Building on the significant potential of the World Wide Web to provide rich and

([1]) Cedefop organised the first workshop to discuss the feasibility of producing pan-European skills forecasts in 2005.

([2]) International Standard Classification of Occupations 2008 sub-major (two-digit) occupation groups and broad levels of formal qualification: high, medium and low.

almost real-time information on employers' needs and job requirements, in 2015–2016, the agency launched a study to assess the feasibility of developing its multilingual system for collecting and analysing data from online job advertisements (OJAs) ([3]). The positive results of this exercise attracted the attention of various key stakeholders, including the official statistics community. Data from this feasibility study were analysed during the first European Big Data Hackathon, organised by Eurostat, the statistical office of the European Union, in 2017.

This success motivated Cedefop to pursue this activity further. In 2017, the agency launched a follow-up project to develop its fully-fledged data production system (DPS) to collect and process OJAs for all EU Member States plus the United Kingdom. Cedefop's system had been developed and was fully operational by the end of 2020. The launch of Skills-OVATE was the culmination of the effort – a central tool for navigating the data. In 2023, the DPS was expanded to cover an additional four countries, and work is under way to broaden the scope of the system by including new sources and variables.

Implementing the agreements in the Scheveningen memorandum concerning the use of big data in European official statistics, in 2016, Eurostat launched the ESSnet big data project, run by a European Statistical System (ESS) consortium of national statistical institutes (NSIs), to incorporate big data in official statistics. The ESSnet big data project included a series of pilots, including one exploring using OJAs for statistical purposes, in which a group of nine NSIs assessed the feasibility of using OJA data in producing official statistics. The group worked on data access and collection issues. The other aim was to examine various methodological aspects, such as aligning with the existing definition of a job advertisement and job vacancy statistics, defining a conceptual model for the target population and extracting information from the advertisement text in natural language. NSI teams also identified different possible statistical products.

At the time of finalising the ESSnet big data project, there was a call to refocus on implementing the most successful pilots to produce statistics. Answering this call, ESSnet Big Data II was launched at the end of 2018 alongside new big data pilots and work on trusted smart statistics for collecting and analysing OJA dataL.

Following up on the 2017 European Big Data Hackathon, where the OJA data collected in the context of Cedefop's feasibility study were used, Eurostat and Cedefop organised a workshop where several Hackathon teams discussed their prototypes with each other and with the team developing the OJA system for

([3]) Project AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14: Real-time labour market information on skill requirements – Feasibility study and working prototype.

Cedefop. The results of that discussion then served as input to developing the DPS within the scope of the 2017-20 project.

The ESSnet Big Data team collaborated closely with Cedefop and the 2017-20 project development team, particularly in the organisation of joint workshops, where developments in the two activities were shared, and the OJA data from Cedefop were shared with the ESSnet. To this end, a partnership agreement was established between Cedefop and Eurostat to coordinate activities and share knowledge. In its final report, the ESSnet Big Data team recommended using the data collected by the Cedefop OJA data system to produce official statistics.

The development of the OJA DPS was not considered a one-off exercise. Therefore, Cedefop and Eurostat made the best use of the investments made during the previous projects to further utilise, maintain and develop the system to secure continuous production of skills intelligence from the OJA data (see also Descy et al., 2019). The joint endeavour led to integrating the Cedefop OJA DPS into the European Web Intelligence Hub (WIH). As of 2020, the system has been developed jointly, utilising both organisations' resources and expertise. Ensuring the high quality of data has positioned OJA-based analyses among the most relevant sources of labour market information. Combining OJA analyses with conventional sources, such as household and employee surveys and skills forecasts, provides a comprehensive view of trends in skills demand and supply in Europe.

This publication maps a decade of Cedefop's work on OJAs. It follows the initial work published in Cedefop (2019a) and reflects the progress made in understanding the strengths and weaknesses of OJA data. Over the years, we have moved beyond the initial stages of data gathering and counting job postings and the frequency with which skills are mentioned. Today, we can conduct detailed analyses of individual job types and specific skills, providing more granular insights for labour market intelligence. These advances have significantly enhanced the quality and reliability of OJA data, allowing a deeper understanding of labour market trends, skills demands and occupational shifts. In this publication, we consolidate these achievements and outline the next steps for further strengthening the use of OJA data within the WIH.

The publication provides an overview of advances in utilising OJA data for labour market intelligence. Chapter 2 presents a landscaping exercise, which is essential for understanding the OJA market and establishing a framework for interpreting the results of analyses. Chapter 3 describes the approach to gathering and analysing OJA data, detailing the individual stages of the DPS and the methodologies used to measure and enhance data quality. Chapter 4 delves deeper into the extraction of skills and occupations, highlighting improvements

made to the system. Chapter 5 further refines the classification of professions and skills, specifically focusing on digital skills, while Chapter 6 examines the role of OJA data in tracking the green transition. The possibility to extend the scope beyond skills and occupations by exploring the classification of education fields, enhancing the understanding of qualifications and their relevance in the labour market, is described in Chapter 7. Finally, Chapter 8 focuses on the overarching goal of this work – delivering robust skills intelligence derived from web data, particularly OJAs, to support evidence-based policy decisions and labour market analysis.

Chapter 2.

Understanding the landscape of online job advertisements

The primary objective of an OJA is to attract the best possible candidate to the vacant post. Thus, the information provided is not necessarily meant to form the basis for labour market analysis or production of official statistics. Yet, we believe that the richness of the data is highly valuable for understanding labour market trends. Therefore, gaining a thorough understanding of OJAs, as the object of the research, through a landscaping exercise is key to drawing accurate conclusions.

The landscaping exercise is a crucial step in producing high-quality data, providing in-depth knowledge of OJA markets across Europe. The analysis encompasses various aspects, from the operational frameworks and business models of OJA providers to the broader economic and regulatory contexts in which they operate. Furthermore, a comprehensive record of each portal's characteristics, as they appear to job seekers and employers, was compiled. The activities build on Cedefop's first landscaping exercise (Cedefop, 2019b), ensuring continuity and incorporating methodological advances over time. In 2021, this exercise was further refined, integrating lessons learned from previous exercises.

To enable a rigorous and systematic selection process, a structured evaluation framework was applied. This approach integrates multiple layers of analysis, incorporating qualitative assessments of a source's significance in the OJA market alongside quantitative measures evaluating its alignment with predefined quality standards. These assessments determine a source's suitability for data production, ensuring that only those meeting the highest quality criteria are considered for inclusion in the system. The final composite indicator, derived from both qualitative and quantitative evaluations, supports prioritisation by identifying which sources should be included first and which do not meet the necessary thresholds for inclusion.

The involvement of individual country experts (ICEs) was essential at this stage. Their expertise in language and knowledge of local labour market dynamics significantly enhanced the quality and accuracy of the analysis. This chapter details the methodology employed for the landscaping activity, outlining the steps involved in source identification, assessment and selection. Furthermore, it explains how the country-specific OJA market analysis was conducted and describes the integrated qualitative and quantitative assessment framework used to evaluate sources. This structured approach ensured the selection of the most relevant

sources, guiding prioritisation for data ingestion and processing phases while upholding consistency and high standards in data quality.

2.1. Identification of OJA web sources

This activity aimed to identify websites that advertise jobs and compile a list of sources with detailed descriptions of their characteristics and OJA features. The approach varied depending on whether the countries were conducting the landscaping exercise for the first time or had already been incorporated into the OJA DPS.

(a) European Free Trade Association (EFTA) countries (Iceland, Liechtenstein, Norway and Switzerland). These countries were conducting the landscaping exercise for the first time.
(b) The 27 Member States of the EU (EU-27) and the United Kingdom. These countries were already covered by the DPS and focused on identifying new sources.

The methodology involved the following steps.

(a) Translating keywords. A predefined list of keywords in English (e.g. 'job search', 'job offers', 'online job search sites') was translated into national languages by the ICEs to ensure consistency in search queries.
(b) Creating a list of job portals. The ICEs conducted anonymised Google searches using these keywords and recorded all resulting job portals.
(c) Assessing the novelty of sources (only for DPS countries). ICEs verified whether each website was a new source, an existing source or a site that was not relevant (e.g. spam, training course advertisements or CV writing services).
(d) Evaluating new sources. The ICEs performed an in-depth analysis of newly identified sources to assess their quality, guided by standardised protocols to ensure consistency.

2.2. Country-specific online job advertisement market analysis

Source identification was complemented by an analysis of country-specific labour markets and the OJA landscape. Each ICE was tasked with drafting or updating a landscaping report using a standardised template, ensuring uniformity in analysis. These reports provided the following insights.

(a) Labour market structure and digitalisation. This included employment trends, sectoral distributions and hiring patterns. It was supported by metrics such as internet usage, digital skills and business digitalisation levels (benchmarked against the Digital Economy and Society Index – DESI).
(b) Legislation, policies and OJA market analysis. This examined legislation and policies that regulate OJAs, labour market reforms and active employment policies. The market analysis covered the number and type of sources, market concentration, key operators and prevailing market trends.
(c) Role of public employment services (PESs) and other recruitment channels. This analysed PES portal coverage and their role in the recruitment ecosystem, along with alternative hiring channels such as social media and corporate websites.
(d) Trends, challenges and specific issues. This covered aspects such as market growth, the impact of the COVID-19 pandemic, multinational operators and online hiring preferences.

2.2.1. Validation of results

The accuracy of the findings was verified through consultations with ICEs, recruiters and web portal owners. Two types of landscaping reports were produced.

(a) For new countries (EFTA), reports followed a structured approach based on previous landscaping exercises. Each ICE received detailed guidance, including on desk research, data analysis and specific sections to be updated. The reports used data from sources such as the Labour Force Survey, Eurostat's information society indicators and the Digital Economy and Society Index (DESI).
(b) For existing DPS countries, reports were updates of previous exercises, incorporating newly identified sources and addressing inconsistencies. ICEs received documentation similar to that provided to EFTA countries, along with the latest OJA data from the DPS. ICEs were responsible for validating these data and providing qualitative insights based on the updated information.

2.1. Source assessment and selection

The landscaping activities outlined in the previous sections generated extensive insights into the OJA market across Europe, covering platform characteristics, business models, market structure and operational context. While this information was comprehensive, it did not provide explicit guidance on prioritising sources for data production. To address this, a ranking model was developed, integrating two

key components: the analytical hierarchy process (AHP) score and the ICE ranking.

The AHP score quantitatively measures source quality based on a structured evaluation of their desirable characteristics. These characteristics are weighted according to stakeholder preferences, allowing a comparative analysis of sources. Using a pairwise comparison method, the model produces a numerical score that consolidates stakeholder input into an overall ranking.

The ICE ranking complements this by providing a qualitative assessment of a source's relevance in the national OJA market. Experts evaluate sources based on factors such as their popularity in online searches, consistency in providing high-quality data and breadth of occupational, regional and sectoral coverage. The ICEs' local expertise ensures that factors specific to each country's job market are considered.

By combining these two measures, the ranking model offers a comprehensive evaluation, enabling the prioritisation of sources for integration into the system. This structured approach ensures that only the most reliable and relevant sources are selected for OJA data production.

2.1.1. Quantitative assessment of websites' characteristics (AHP score)

The AHP is a structured method used for multi-criteria decision-making, particularly effective when multiple factors must be evaluated simultaneously. It breaks down decision criteria into a hierarchical structure, allowing systematic comparisons between elements. The method combines preferences expressed by multiple stakeholders into a single output. The criteria to which the method was applied in this case was the subset of all the source characteristics assessed by the ICEs in the previous step of this exercise. This enabled a comprehensive assessment of website characteristics (e.g. Google ranking, type of source) and completeness of job postings (e.g. inclusion of occupation and publication date). An online tool was used for this purpose: AHP-OS. Stakeholders provide input by comparing criteria in pairs, reflecting their relative importance. Once priority weights are determined, they are mapped to the observed characteristics of each source. The final AHP score is computed as the algebraic sum of these weighted values, offering a numerical ranking that integrates stakeholder preferences into a structured evaluation process.

Using the AHP model ensured transparency and allowed multiple stakeholders to participate. It provided a robust framework for evaluating a large number of criteria and ensured that the sources selected met high standards of data quality and relevance.

2.1.2. Qualitative assessment of website relevance (ICE ranking)

The qualitative assessment of website relevance was based on some structural and technical characteristics of the source that were not obvious by inspecting the sources directly but could be obtained by analysing the downloaded data. ICEs were involved in evaluating the importance of each source, considering the evidence provided and their knowledge of the labour market in their country. This evaluation measured the popularity, stability and coverage of each source.

A popularity study was conducted for countries that had started the landscaping process (EFTA countries) and those that had updated existing data (EU-27 and the UK). It involved measuring individual sources' popularity by assessing the frequency of web searches that referred to them. This was done through Google Trends, which is provided centrally and produces an index of search interest based on the volume of search normalised by region and time range.

The stability and coverage were assessed only for the EU-27 and the UK and were based on historical data from sources collected by OJA DPS.

Stability was evaluated by examining a source's performance over time. This assessment looked at the consistency of OJA availability and the presence of outliers in the data, which helped evaluate the source's reliability and trustworthiness.

Coverage was defined as the source's ability to cover all a country's occupations. From a data quality perspective, this can be seen as a proxy for the source's completeness. Each country's source coverage was evaluated based on the distribution of OJAs classified according to the first occupation digit in the 2008 International Standard Classification of Occupations 2008 (ISCO-08), using the Eurostat Labour Force Survey data as a reference for the country's employment structure by occupation. The source occupation distribution and data from the Labour Force Survey were compared at the country level and an assessment was made on whether – and to what extent – the list of sources could replicate the Labour Force Survey benchmark, indicating the representativeness and comprehensiveness of the sources.

2.1.3. Final evaluation and decision rule

The final ranking model integrated both AHP scores and ICE rankings to establish a combined priority score. To ensure consistency, sources were grouped into quartiles, with those in the highest ranking quartiles prioritised for inclusion in DPS development. This approach ensured that sources were evaluated using a combination of quantitative and qualitative measures, enhancing the robustness of the selection process and the overall reliability of the OJA DPS.

For existing DPS countries, only sources in the top two priority groups were selected. However, for new countries, a more flexible approach was used. Switzerland followed the same selection criteria as DPS countries, Liechtenstein included an additional priority group and Iceland and Norway considered all potential sources to accelerate their integration.

This structured evaluation and selection process ensured that the sources integrated into the OJA DPS were of the highest quality and relevance.

2.2. Concluding remarks

Between 2017 and 2021, European labour markets experienced significant changes due to the long-term digital transformation and the COVID-19 pandemic. The adoption of digital technologies accelerated, and online recruitment channels expanded.

The following key trends can be observed.

(a) Growth in the OJA market. The number of job advertisement sources has increased, with multinational operators expanding while national portals have remained stable.
(b) Transformation of PES portals. PES platforms have evolved, incorporating AI-driven job-matching tools and serving as aggregators of job listings.
(c) Market consolidation. The OJA market has shown signs of increasing concentration, with larger platforms gaining market share.

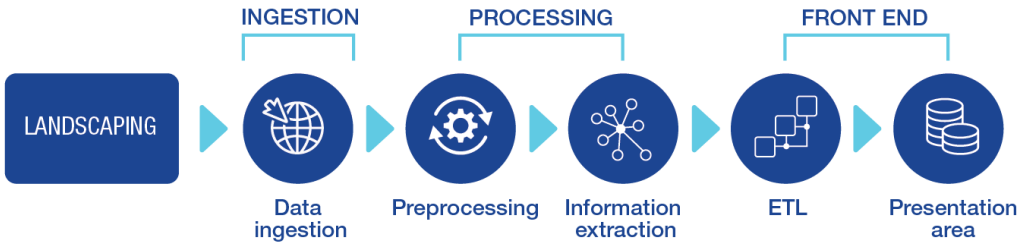
Chapter 3.

Gathering and analysing online job advertisements through the data production system

The gathering and analysis of the OJAs are organised through a multilingual modular DPS (see Figure 1; for detailed information on the hardware and software components, see Annex 1) ([4]). The OJA web sources identified throughout the landscaping exercise (see Chapter 2 for details) enter the 'data ingestion' phase. In this phase, the system downloads the content of each individual OJA. The downloaded documents are cleaned of irrelevant content during the next phase – 'preprocessing'. The 'information extraction' phase leads to the creation of structured data (sector, occupation, geographical unit, etc.) describing the content of the OJAs.

The content of OJAs is then stored in various databases and used for research purposes, labour market intelligence and the production of statistics and indicators. The following sections present a more detailed description of the stages of the DPS.

Figure 1. **Data production system**



ETL = extract, transform, load.
Source: Cedefop (2019a).

3.1. Data ingestion

All sources suggested as an outcome of the landscaping exercise (see Chapter 3) enter the data ingestion process. For each source, we use individual, tailor-made ingestion solutions, depending on whether the data are gathered from the front end

([4]) This chapter builds on an earlier, detailed description (Cedefop, 2019a).

(the part of a website visible to users) and/or the back end (databases and systems powering the website, which its operator can provide access to). The individual solutions are based on the following methods.

(a) Scraping is used to extract structured data from websites. Web scraping implies that the data are already structured on the web page and can be extracted precisely by knowing the exact position of each field (such as the OJA title, contract type offered or skills required). As specific web scrapers need to be developed independently for each website, they are ideal for websites with many vacancies.

(b) Crawling uses a programmed robot that browses web portals systematically and downloads their pages. It is much more generic than scraping and more straightforward to develop. However, crawlers collect much more website noise (irrelevant content), and more effort is needed to clean the data before further processing.

(c) Direct access via an application programming interface (API) enables downloading of OJA content directly from portal databases. Accessing web content via an API often requires a formal agreement with the website operators and incurs some maintenance and agreement costs. These agreements establish retrieval conditions, scheduling preferences and alternative data access channels (e.g. API or file transfer). The WIH prioritises transparency and collaboration with website owners to minimise the burden and ensure adherence to standard internet conventions. API-based access is governed by WIH agreements on web sources and data access guidelines, aligning with EU statistical regulations such as Regulation (EC) No 223/2009, which protects confidential data while permitting its use for statistical purposes. These rules and conditions are currently under approval and consultation to ensure alignment with the recent revision of this regulation.

3.2. Data preprocessing

As the primary objective of an OJA is to attract the most suitable candidate for an advertised post rather than to serve as a source of labour market information, sources vary in quality, content, detail and structure. Therefore, before the information is extracted from the text, the documents must be transferred to a format suitable for this activity. This preprocessing involves the following.

(a) Cleaning. In addition to analytically useful information, OJAs contain various ‘noisy’ elements (such as unrelated advertisements, unticked options in drop-down menus, company profile presentations). Cleaning is a sequence of

activities to remove such noise from the data and prepare them for the following phases.

(b) Merging. Employers often post an OJA on more than one portal. OJA website aggregators increase the chance of finding duplicates, which are undesirable to have in the final analysis. However, duplicates can enrich the data in the initial part of the preprocessing phase, as some portals might contain additional information for the same job posting.

(c) Deduplicating. Once the data for the same vacancy have been merged, duplicate vacancies must be removed from the analysis. An OJA is considered a duplicate if the description and job location are the same as in another job advertisement in the database. In addition, vacancy metadata (such as reference ID and page URL) are used to identify and remove job vacancy duplicates on aggregator websites.

3.3. Information extraction

To extract the information, we use a language-based rather than a country-based approach. This means that the OJA is processed in its original language regardless of the country (or portal) it was identified in. Therefore, the information extraction process starts by identifying the language ([5]). Each OJA provides a rich source of information.

(a) Regularly produced variables are extracted, updated and published in various data outlets. Examples of regularly produced data are job title (according to ISCO), job location (including information on the region using the nomenclature of territorial units for statistics), economic sector (using NACE) and skills (using the European Skills, Competences and Occupations (ESCO) classification system).

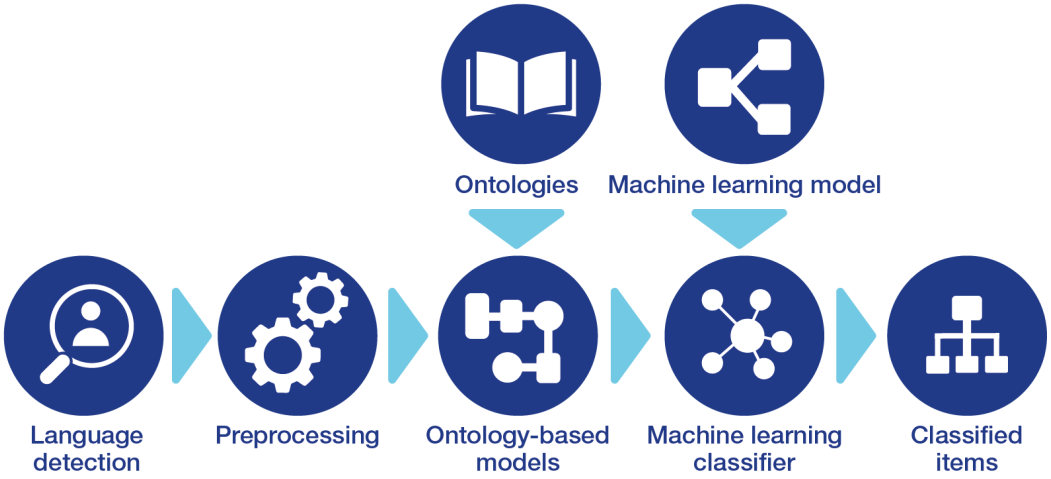
(b) Experimental variables are information that is regularly produced but not (yet) suitable for public presentation for various reasons, such as low yield, incompleteness of information or insufficient level of testing (type of contract, experience, wages).

(c) Other variables are those contained in the OJA but not yet tested for their suitability for production or variables tested for production, but the value they add to the dataset is inadequate considering the resources needed for production (e.g. fields of study according to the International Standard Classification of Education (ISCED)).

([5]) The DPS can recognise and process all official EU languages and other languages that are widely used in some countries, such as Basque, Catalan, Gaelic and Russian.

Each variable is extracted through an individual language-specific pipeline using two essential features: ontologies and machine learning models (Figure 2). Large language models (LLMs) are also piloted on a few variables. Some initial results indicate a positive effect on the precision of data extraction, especially on their ability to classify data more granularly. However, this is very often associated with higher and less predictable production costs and questions about the consistency of the results. The other issue in the multilingual system is the comparability across the various EU languages. Therefore, more groundwork must be done before using LLMs for overall data production.

Figure 2. **Process of information extraction**



Source: Cedefop (2019a).

Ontologies create a framework for processing and analysing OJAs. The DPS uses standard and customised ontologies related to the skills and jobs market. With the power and flexibility of machine learning algorithms, the content of job advertisements is matched to the available ontology terms, such as occupation, industry, region of the workplace and type of contract. The process first tries to classify an OJA based on text matching and similarity with the relevant ontology (such as matching vacancy titles with ISCO occupation titles). If no result is obtained, a machine learning algorithm ([6]) (classifier) decides on the classification.

The classification accuracy is regularly monitored, and all ontologies are continuously updated and enriched. The results of the classification process are

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

periodically validated by experts (for details, see Section 3.4). The outcomes of this process (proposed corrections) are used to improve the accuracy of the classifier. The semi-automatic augmentation process adds new terms and synonyms not yet included in an ontology through machine learning algorithms, which are approved by expert checkers. Ontologies can also be updated manually to reflect new information (new occupation trends or even updating the underlying ontology such as ESCO).

3.4. Monitoring and improving data quality

OJAs are a source of high-frequency, quickly produced and granular data. However, one must be aware of limitations in terms of representativeness and treat them with appropriate caution (e.g. Napierala, Kvetan & Branka, 2022). Key considerations regarding the reliability of vacancy data include the following.

(a) OJAs represent only part of job demand; not all job vacancies are advertised online, and some jobs are more likely to be advertised online than others. It can be expected that the OJA data are subject to occupational or qualification bias.
(b) In most countries, the OJA market comprises multiple actors with different business models. There is usually no overarching source for all OJAs. Therefore, the volume, variety and quality of the data depend on the portals selected for analysis.
(c) Penetration of OJA markets varies in and across countries and changes over time. Low internet penetration and lack of basic digital skills across the population are key factors influencing employers' decisions on the extent to which they use OJA portals as a recruitment channel.
(d) OJAs and the information extracted from them must be processed by the various tools described above to produce viable data. Even the most up-to-date techniques are still subject to error.
(e) Finally, all ontologies are developed to sort and organise a diverse and complex universe. Despite the enormous effort invested in creating them, they are still imperfect and may contain systemic errors that can be corrected only over time.

A detailed and robust quality assurance process has been designed and adopted to tackle these challenges. Tools and checks are regularly used to mitigate potential data biases. Documentation of the process, suggestions, revisions, ontologies and training sets are made available under Creative Commons licences to enable potential contributors to evaluate and improve them.

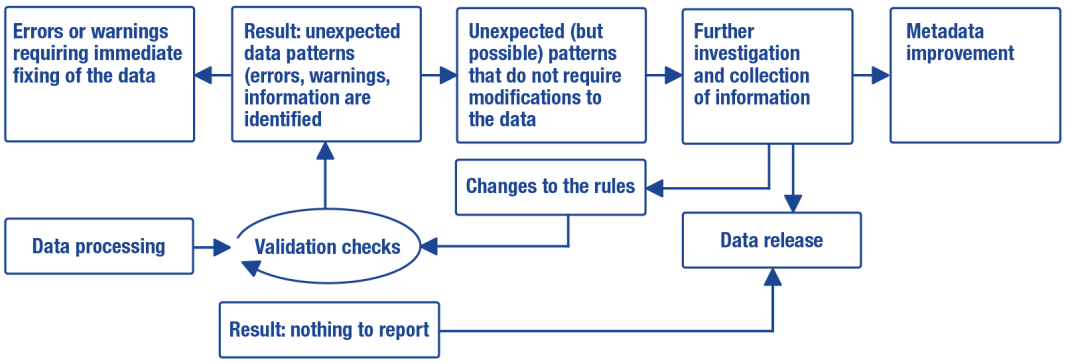
3.4.1. Data validation process

Data validation is critical to objectively identify data errors and abnormal patterns. Despite not directly measuring the overall quality of the data, it helps us improve dictionaries, ontologies and classifiers and fine-tune the classification and machine learning processes. The outcomes of the validation process are also used as an input to the machine learning process (e.g. by adding terms to dictionaries to map educational titles correctly or by defining wrong associations between occupations and skills). This ensures ongoing development and improvement of the system.

Validation is executed manually over smaller datasets (about 1 000 per language pipeline). Afterwards, the results are translated into model improvements or ontology enrichment. This iterative process (execution – validation – fixing – execution) can be repeated until an appropriate level of quality is reached. More details about the validation process and quality monitoring are given in Section 3.4.2 and Figure 3.

The validation process is designed to ensure the accuracy and reliability of the OJA datasets produced. It involves using validation rules to check data compliance, covering aspects like consistency with official code lists, hierarchical classifications and distribution stability over time ([7]). When anomalies are detected, WIH statisticians revise the data, gather information on anomalies and modify rules.

Figure 3. **Organisation of data validation process**



Source: Authors.

The validation cycle is an ongoing dynamic process that allows continuous improvement and adaptation of the rules. Rules refer to predefined conditions and logical constraints designed to ensure data quality, consistency and integrity. They serve as a framework for identifying errors, enforcing standardisation and facilitating the systematic validation of data.

([7]) For more information, please see Eurostat's methodology for data validation.

Rules encompass various aspects, including consistency within classifications, absence of missing data for certain variables and adherence to formatting requirements for string variables. These rules can be constraints (e.g. specific variable types, required fields) or guidelines (e.g. recommended formats, best practices for consistency).

Structural validation rules ensure the overall integrity of the database, checking for correct naming and the absence of empty fields. They emphasise continuous improvement and adherence to the requirements of official statistics. Therefore, while the validation process is fundamental for data improvement and quality control, its rule set is always amenable to change. The adaptability of rules ensures that validation mechanisms remain responsive to new challenges, regulatory updates and technological advances, ultimately strengthening the robustness of the data validation framework.

3.4.2. Quality monitoring

Quality monitoring is another crucial step in developing a system for generating reliable and robust data. While the validation process is part of regular data production, quality monitoring represents a set of one-off exercises developed outside the DPS. Various procedures are being developed and put in place to evaluate and assess the algorithms used to harvest web data sources collected by the WIH. These procedures combine human and machine intelligence to maximise accuracy and use machine learning to assist in human tasks to increase the efficiency of the classifiers.

Creating a gold standard for OJA classification represents one way of addressing this need for evaluation and quality improvement of algorithms (Nagy & Reiss, 2023; Nagy et al., 2024). A cyclical evaluation process is intended to initiate iterative feedback and enhancement cycles. Annotators will examine specific advertisements within the evaluation dataset, indicating their agreement or disagreement with the classification outcomes through a data annotation tool. Additionally, they have the option to suggest modifications to the algorithms to address issues identified. The evaluation dataset will result in:

(a) evaluation metrics for the classification algorithms (e.g. the accuracy rate);
(b) suggestions for improvement of the sets of keywords (i.e. ontologies);
(c) a set of human-labelled data growing over time for training machine learning models.

3.5. Data access

Data access is organised through the WIH across different domains (layers) based on user roles and the nature of the data. The WIH domains ensure controlled access to data, considering the sensitivity of the content and the specific purposes for which it is accessed.

The data are available via the WIH DataLab, a data analysis environment established by Eurostat to explore and analyse web data. It is built inside the web intelligence platform and features a big data cluster capable of processing large datasets. Jupyter Notebook and RStudio Server power it. The DataLab focuses explicitly on data obtained from OJAs across all countries covered by DPS.

The data are processed with information organised by individual classified variables (occupation, skills and education level are only few). The data are still in the experimental stage, and results derived from the DataLab may be biased. Cedefop and Eurostat are actively working to enhance the data quality, aiming to establish the lab as a reliable resource that meets high standards in the future.

Eurostat ensures timely and efficient data dissemination to relevant stakeholders. Data from specific data flows are accessible to restricted users every quarter with a one-month delay. The data release is accompanied by a WIH blog post, data release notes and a validation report explaining the main data updates and improvements made over the past quarter.

Access to the restricted domain is provided on request. Requests can be submitted to the email address given on the Eurostat web page dedicated to the WIH. Users receive access to the WIH wiki space, a collaborative space available only to a specific community of users, that provides users with the latest and most valuable information on developments and the progress made by the WIH (access granted on request) on exploring and using web data to produce official statistics.

3.6. Collaboration and coordination

To boost awareness of OJA data across Member States and to involve NSIs in data production, Eurostat initiated the creation of the Web Intelligence Network (WIN), a community centred around the WIH to advance statistical methods and tools related to using web data in official statistics.

Specifically, the WIN assists the ESS in using tools and technologies for web data collection and processing, provides user support and documentation for the WIH and contributes to the improvement of platform components. It also plays a role in integrating WIH services into national statistical systems, exploring new web data sources, assessing data quality and producing experimental statistics. Methodological developments include creating quality frameworks, refining methodologies and developing correction methods for web data sources. Additionally, the network is tasked with producing methodological handbooks and guidelines on official statistics. The network actively disseminates knowledge through conferences, training activities and collaboration with international organisations.

3.7. Concluding remarks

Over the last 10 years, Cedefop and Eurostat have established a comprehensive framework for systematically collecting and analysing OJAs through a multilingual, modular DPS. This system leverages various data-gathering methods (scraping, crawling and API access), each tailored to the unique demands of individual OJA sources. The data ingestion process, followed by rigorous preprocessing steps, ensures that only relevant, deduplicated and clean data go forward for analysis, minimising noise and redundancy. The DPS's information extraction phase takes a language-centred approach, employing ontologies and machine learning models to classify and structure job advertisement data.

Data quality is a paramount concern for the DPS, addressed through structured validation and quality monitoring practices. Regular validation processes support the refinement of dictionaries, ontologies and classifiers to enhance accuracy, while a cyclical approach to quality monitoring, incorporating human input, further strengthens the reliability of extracted data. Establishing evaluation datasets and integrating a gold standard for OJA classification demonstrate an iterative commitment to continuous quality improvement.

Chapter 4.

Extracting skills and occupations

The ultimate goal of the process organised within the DPS is to extract information from OJAs. Although the DPS classifies OJAs by sector, region, education level and many other variables, this section describes only the extraction/classification of information on skills and occupation (full details of the basic characteristics of the DPS are provided in Cedefop (2019a)). These two variables are vital for further combination with other 'conventional' data sources (such as surveys and/or forecasts) to guide vocational education and training policies and practices.

The DPS employs a blend of machine learning and ontology-based techniques for information extraction to ensure that the final results are robust and accurate. While machine learning techniques offer better and more flexible results, the ontology-based approach ([8]) helps to address the DPS's multilingualism and provide more stable and comparable data extractions across time.

The WIH-OJA system encounters two primary challenges regarding these two variables:

(a) identifying occupations and skills as listed in OJAs with the pre-existing categories given in the relevant taxonomies (for more details, see Section 4.1);

(b) detecting and cataloguing emerging occupations and skills as the labour market evolves (for more details, see Section 4.2).

4.1. Identifying occupations from online job advertisements

The occupations are extracted and classified based on ISCO-08 ([9]), assigning labels to OJAs at the unit group level (ISCO four-digit), which encompasses a spectrum of over 400 distinct occupation titles. However, given the intervals between the updates (the latest version was updated in 2008), ISCO struggles to capture emerging occupations that did not exist more than 15 years ago. Therefore, even at its most detailed level, ISCO is quite general and often uses groupings that say little about the nature of the jobs included.

([8]) For this purpose, the DPS uses ESCO v1.1.1, ISCO-08 and O*NET as overarching taxonomies for skills and occupations.

([9]) International Standard Classification of Occupations.

In many ISCO-08 occupations, the title ends with 'not elsewhere classified', which often bundles together many distinct jobs, the number of which can be rather large (a few of these groups represent over 10% of OJAs). Moreover, the groups can contain very heterogeneous jobs or job titles. For example, the ISCO unit group 3119 – 'physical and engineering science technicians not elsewhere classified' – includes aviation safety officers, footwear product developers, offshore renewable energy technicians and quality engineering technicians, which are four very distinct jobs with diverse skill sets. Another challenge of classifying job titles into unit group occupations lies in the classification process. All classifiers rely on the job titles to perform the occupation classification. Therefore, the process is strongly affected by the complexity of individual languages and how employers phrase/assign job titles.

4.2. Using large language models to classify occupations

To address the issues flagged above, attempts were made to classify occupations according to the ESCO individual occupations (one level below the four-digit unit group ISCO code). This meant classifying the OJA titles into approximately 1 500 occupation names. Moreover, the technological advances in big data computing and machine learning experienced throughout the project's development, especially the widespread use of LLMs, allowed us to develop new occupation classification methods. The pilot exercise was carried out on a sample of OJAs in five countries: Finland, Germany, Italy, Spain and the United Kingdom. The countries were selected based on the overall size of their labour markets and coverage of OJAs. Finland was chosen mainly due to the complexity of the language, as a result of which ontology-based extraction faced specific difficulties.

Although the use of LLMs could yield better results in the classification of job titles, there is a clear issue with resources in terms of volume and predictability. Initial testing confirmed that using a 'direct approach' (i.e. asking an LLM to assign an adequate ISCO code to the job title as listed in an OJA), would take about three seconds for each OJA contained in the DPS. Considering the scale of the task (with hundreds of millions of OJAs), the total time taken would be impractically long and the resources required difficult to predict.

Therefore, an alternative strategy has been developed based on independent embeddings (processing OJA texts into dense vector representations that preserve the semantics of the texts) of OJAs and occupations. This approach involves calculating embeddings for each OJA and each occupation only once. The embeddings for the OJAs and the occupations were pre-computed and stored.

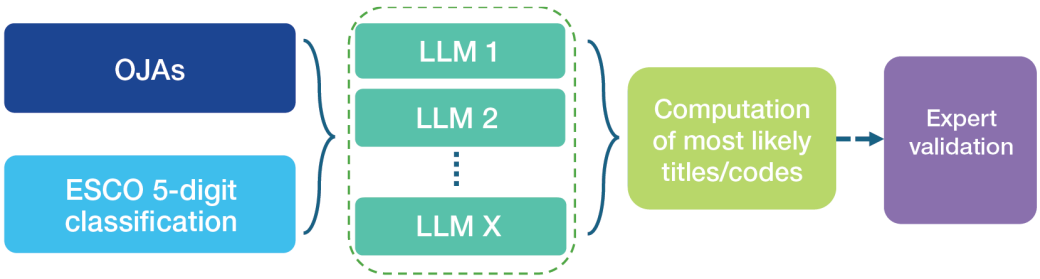
When classifying, the system compared the OJA embeddings with the occupation embeddings to find the best match.

As the market currently offers various LLMs that differ in stability, various LLMs were tested and deployed in the initial classification. This helped us to reduce the model-specific biases and improve the overall reliability of the dataset. The process involved the top-tier LLMs based on the 'massive text embedding benchmark' (Muennighoff et al., 2023). These models are known for their high level of performance in generating embeddings from textual data. The selected models generated embeddings for all OJAs for each country.

Starting with every ESCO occupation (five-digit level), each model was utilised to identify the most relevant OJA contained in the DPS. This involved matching the occupation's characteristics (alternative labels, description and skills) with the OJA embeddings. An agreement computation regarding prediction scores between the models was also conducted. This step was crucial for filtering out occupations not represented well in the corpus (with a low similarity score), such as political figures or military roles, which might not be relevant to the dataset.

Each OJA was associated with three potential labels corresponding to the highest-rated suggestions of the LLM. The following critical step involves human experts validating each association between the OJA and an ESCO five-digit classification. The suggestions generated by LLMs were assessed in a human validation process, ensuring the accuracy and relevance of the classification (Figure 4). At the same time, this maintained a diverse representation of occupations and mitigated the risk of biases or errors from any single model. The OJAs included in the dataset were sourced from real-world scenarios, avoiding the limitations of synthetic data.

Figure 4. **Development of the detailed occupational classification**



Source: Authors.

The exercise yielded the following results (Table 1). The English pipeline produced the most accurate results, ensuring that the OJAs tested were completely covered by valid ESCO occupations (the ESCO classification contains

more than 3 000 occupations). The outcomes for the German, Italian and Spanish pipelines were also very good, equalling an approximately 75-80% success rate (the success rate is considered to be the algorithm's ability to assign precise categories). The Finnish pipeline achieved the worst results, with less than half of the matches correct.

Table 1. **Distribution of OJA five-digit benchmark by country**

Country	No OJAs	One valid	Two valid	Three valid	Title insufficient (*)
UK	2 520	2 256 (89.5%)	349 (13.8%)	129 (5.1%)	13
IT	2 878	1 450 (50.4%)	460 (16.0%)	256 (8.9%)	394
FI	1 897	389 (20.5%)	174 (9.2%)	108 (5.7%)	152
ES	2 024	1 088 (53.8%)	316 (15.6%)	181 (8.9%)	130
DE	2 493	1 339 (53.7%)	294 (11.8%)	198 (7.9%)	107

(*) Expert assessment that, based on the OJA title, it was not possible to classify the occupation.
Source: Authors, based on OJA DPS.

From the perspective of the WIH DPS, it is crucial to assess whether the LLM-based method achieves better accuracy than the currently prevalent ontology or fuzzy matching. At the ISCO four-digit level, the improvement is notable (except in English, where the accuracy achieved using traditional approaches was already very high; see Table 2.). The improvements in the Finnish pipeline classification are the most substantial, bringing the accuracy level closer to the average of other languages. However, the accuracy rates fall slightly at the five-digit occupation level (although they still exceed the success rates of ontology or fuzzy matching at the four-digit level).

Table 2. **Accuracy of LLM-based classification (and comparison against the performance of WIH (four-digit) classification) by country**

Country	WIH 4-igit (%)	LLM 5-digit (%)	LLM 5-digit (%)	MAE	No OJAs
UK	94.04	96.18	94.44	71.1	2 357
IT	69.78	75.35	69.82	382.09	1 882
FI	45.90	63.41	57.52	653.58	645
ES	66.13	80.77	76.62	282.62	1 326
DE	67.14	78.4	74.3	414.47	1 634

MAE = mean absolute error.
Source: Authors' compilation, based on the OJA DPS.

The LLM-based classification system also performs better than broader categorisation models in that it accurately classifies jobs that would have

previously fallen into a catch-all ‘not elsewhere classified’ category. Instead of being lumped together under a generic classification, many of these jobs are recognised for their specialised functions, one of the most desirable outcomes of the whole exercise.

Similarly to the ontology-based approach, the misclassifications naturally also occur in in the LLM-based system. However, they are somewhat easier to detect. The conclusion of the performance evaluation was that the misclassifications usually had a low prediction score, meaning that the system was not very confident about the outcome. Recognising this pattern, it will be easier to flag OJAs where there is likely to be insufficient information to make an accurate classification. The experts validating the LLM exercise concluded that usually between 5-10% of OJAs in the sample fall into that area (and it was over 13% in the case of Italian). English, unsurprisingly, had a negligible share of those (less than 0.5%).

This could lead to a significant quality control measure being established in the classification system. Setting a threshold score allows the system to flag potential mismatches for further review, thus reducing the likelihood of such errors persisting in the dataset. This approach would help further improve the accuracy of automated classification by incorporating a mechanism for identifying and correcting outliers or doubtful classifications or discarding them as unclassifiable. It is, however, essential to note that such a threshold could disproportionately affect occupations rarely appearing in the dataset. This method tested for five languages will be developed for all languages of the DPS in future. However, the entire data production for ESCO occupations will be subject to detailed testing of the precision, costs and final added value.

4.3. Extracting the information on skills from online job advertisements

Skills classification ([10]) relies on ontology matching, utilising ESCO as an overarching taxonomy that encompasses approximately 13 500 foundational terms, augmented by several thousand additional alternative labels for each language variant. It is regularly updated ([11]) to incorporate the latest skills developments and phase out outdated ones. Its structure spans several domains, including skills, knowledge and transversal competences, with a dedicated section for languages.

([10]) For the purpose of this exercise we understand ‘skills’ as a unifying term encompassing knowledge, skills and competence as defined in ESCO v1.1.1.

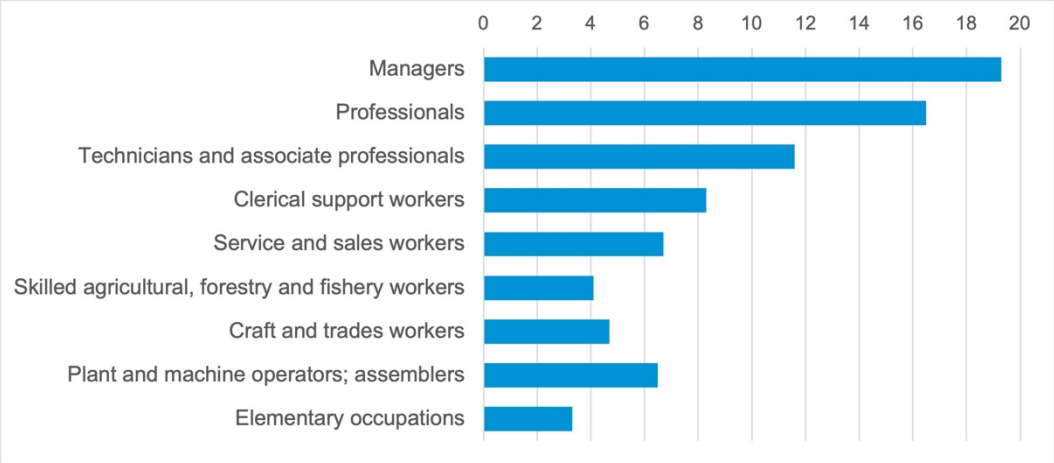
([11]) A more recent version, ESCO v1.2, was launched in May 2024.

The process of skills classification is iterative. Firstly, each ESCO skill's preferred and alternative labels are passed through a word-embedding model to obtain a list of the most similar n-grams, which constitute the skill's markers. These terms are used to match the text in the descriptions of the OJAs so that, every time a marker is found in the text, the corresponding skill is attached to the OJA. The association between markers and skills involves a many-to-many correspondence so that at least one marker points to one skill, while a marker can point to more than one skill.

Extracting skills from OJAs, however, presents considerable complexities. Unlike job titles, which tend to be made explicit in OJA titles, the necessary skills are usually expressed in the unstructured sections of job advertisements. They are articulated in many ways and greatly influenced by linguistic nuances, such as inflexions, translation fidelity and the diversity of terminology. These factors, alongside the variability in employer preferences for describing skills, affect the success rate of the classification exercise.

There are additional challenges to consider. Despite the breadth of ESCO's alternative labels, they may not align with the specific language employers use in job postings. Often, certain skills are not overtly requested, as they are presumed to be inherent to the job role. For example, the skill 'computer use' might go unmentioned in a software developer's job description because it is assumed to be a given. Employers may also rely on stated qualification requirements as proxies for various skills, eschewing the need to list each skill individually. Consequently, the granularity of skills detail varies across professions, with an average professional OJA revealing up to16 skills (Figure 5).

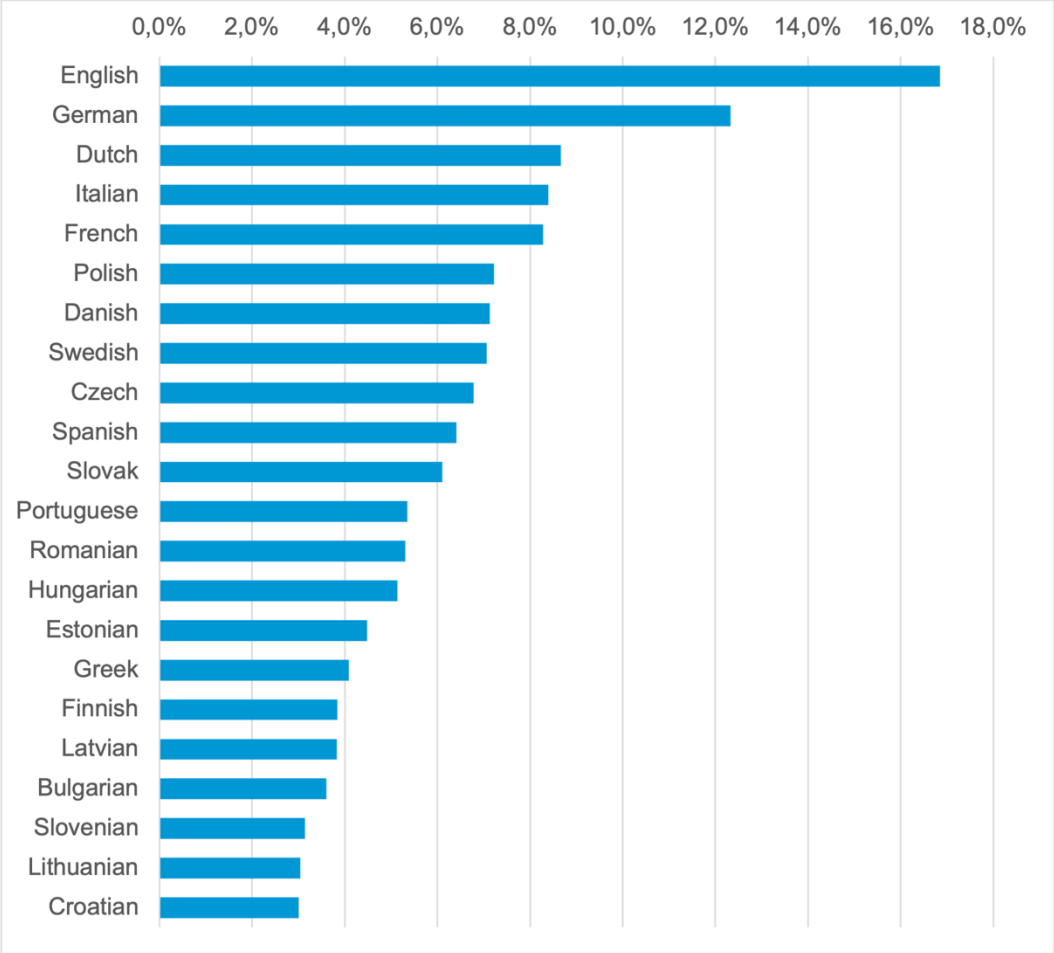
Figure 5. **Average number of skills retrieved from OJAs by occupation, 2023**



Source: WIH-OJA data monitoring system.

When examining the skills yield across occupations or countries, it is evident that the DPS can track only a small part of the ESCO skills. The highest yield is captured in English, followed by German, Dutch, Italian and French, while the lowest yields are found in Croatian, Lithuanian and Slovenian (Figure 6).

Figure 6. **Coverage of skills concepts by language, 2023 (all ESCO terms = 100%)**



Source: WIH-OJA data monitoring system.

The difference in skills yield is driven by three main factors, namely:

(a) insufficient ontology, in that the OJAs do not contain all the alternative terms used by employers or the skills terms used are subject to language specificities (e.g. using verbs rather than nouns when describing skills);

(b) a lack of description of skills in OJAs due to cultural differences (the description of skills can be richer in detail in some languages than in others);

(c) differences in the performance of algorithms in the different language pipelines.

4.4. Augmenting skills ontology

The assessment of skills captured across countries, languages and occupations is a regular activity that is part of the quality monitoring framework. This activity helps us to identify points of interest, such as large and unexplainable disparities in skills capture, and to carry out content-focused validation, which assesses the plausibility of skill–occupation or skill–sector combinations. These points of interest feed in to the design of a methodological approach for improving the extraction of information from OJAs. The approach is also strongly based on data-driven techniques, but its most vital aspect is the extensive involvement of individuals with language and domain expertise. Understanding the differences in skills extraction across languages is a crucial element in ensuring the quality and comparability of results.

However, more detailed activities have been devised to understand the issues in skills extraction described in Section 4.3. The English pipeline provides richer skills extractions for various reasons:

(a) the English language is rather semantically simple compared with other languages;
(b) the ESCO classification was primarily developed in English and then translated into other languages;
(c) most of the natural language processing algorithms have been developed to work in English.

As the WIH DPS processes OJAs in their original language, a set of activities was designed to understand the following:

(a) human-driven augmenting of ontology to increase the number of skills concepts (terms) as defined in the ESCO classification, that are present in OJAs;
(b) computationally driven augmenting of ontology by capturing skills currently not assigned to ESCO skills concepts, which may consist of novel skills and terms or concepts used by employers that are not captured by ESCO alternative labels.

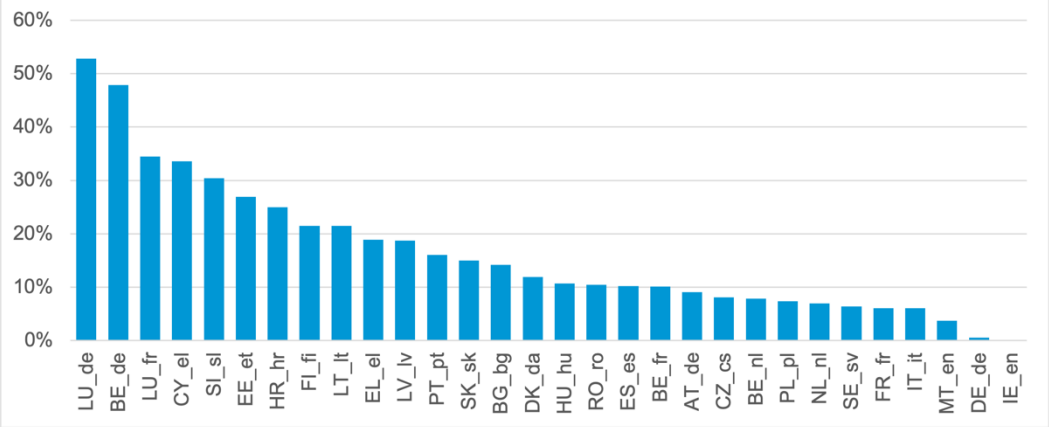
4.4.1. Human-driven augmenting of ontology

The starting assumption of this activity is that, across countries and languages, OJAs are similarly constructed in terms of style and length. This suggests that the WIH DPS would be expected to extract a comparable tally of skills for each language pipeline. If the DPS can identify similar occupations across various languages, we would expect it to capture similar skills. However, in many cases, the skills terms captured for occupation A in English do not correspond to

equivalents in other languages, despite occupation A being identified in other language pipelines (see Figure 7).

Therefore, we used the group of ICEs to examine the first thousand most frequent skills terms (excluding the language skills terms) in the English pipeline. The primary assumption was that, if we observe a relatively equal distribution of occupations across countries, we could expect a similar distribution of skills terms when describing the requirements for these occupations. In many countries, at least a hundred skills terms were missing. We proposed that the ICEs examine 2 932 terms (skills-language combination). For each language, the experts assessed whether the preferred label for the skill, as defined in the ESCO classification, was sufficient and whether the alternative labels for that skill were comprehensive.

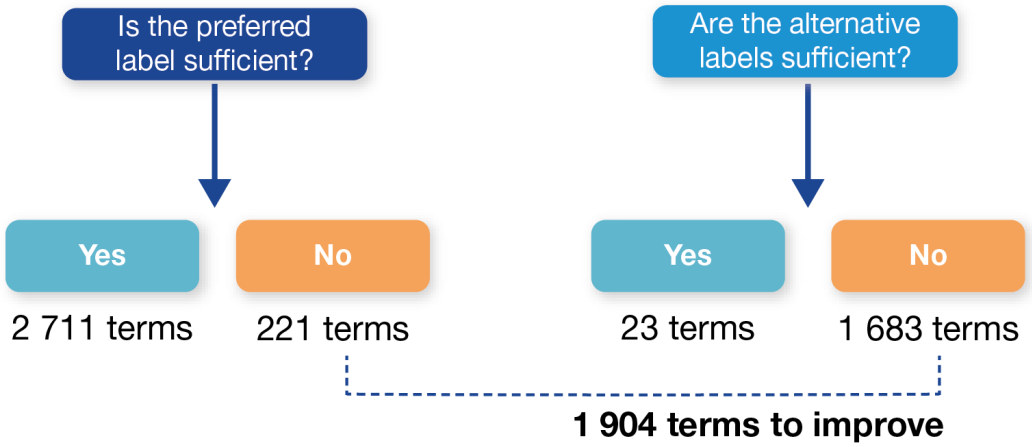
Figure 7. **Share of skills present in the English pipeline but missing in other language pipelines by country and language**



Source: WIH-OJA data monitoring system.

The quality of the preferred labels was generally considered good. Only in less than 10% of cases (221 out of 2 932) were the preferred labels considered insufficient and new terms proposed. In the case of the alternative labels, improvements were proposed in almost 1 700 cases, which, combined with the new preferred labels, accounted for 1 904 language-skill combinations to be improved (Figure 8).

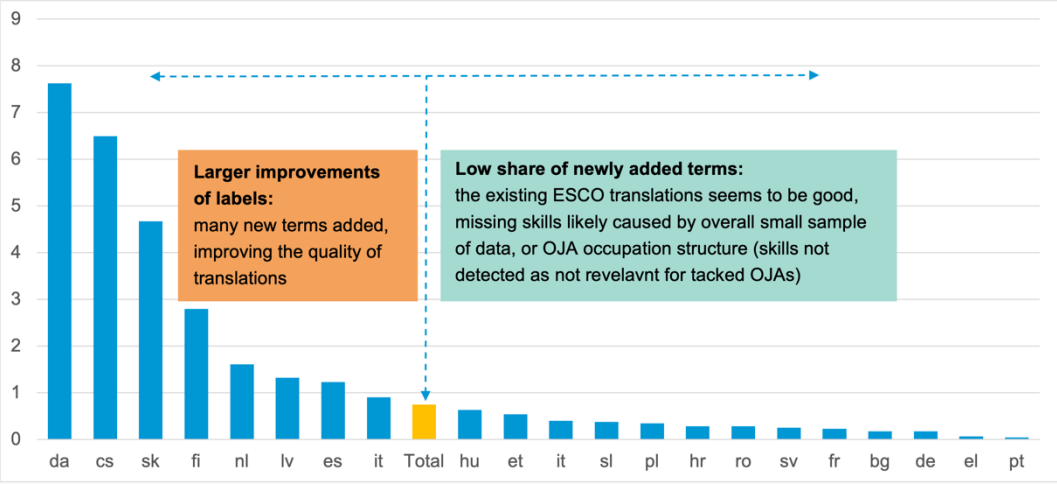
Figure 8. **Responses of ICEs when identifying the skills terms**



Source: Authors, based on OJA DPS.

In many cases, more than one new alternative label was proposed for a total of almost four thousand (3 883) language–skill combinations. Measured as a percentage of the proposed corrections to the skills already detected in a language pipeline, the largest shares of improvements were proposed for the Danish, Czech and Slovak language pipelines (Figure 9).

Figure 9. **New language–skill combinations proposed as a percentage of already detected skills, by language**



Source: WIH-OJA data monitoring system.

In most languages, however, the share of newly added terms in the total number of already detected skills remained below 1%. The skills terms identified in English but missing in other languages could thus be a result of either the overall

small sample size or differences in OJA occupation structure, as the skills not detected are likely to be not relevant for the country's OJAs.

4.4.2. Computationally driven augmenting of ontology

This activity aims to capture skills terms that are not yet part of the vocabulary used by the WIH DPS. Despite being driven by AI methods, this activity relies heavily on human input, since the AI-based proposed terms are validated by country experts against the most similar ESCO skills and the occupation codes in which the terms appeared. The activity was first piloted in the English, French, Italian, Romanian and Spanish pipelines. The approach was recently expanded to cover all 27 Member States and the United Kingdom. The expansion to all original WIH DPS countries also utilises technology advances not fully available during the pilot exercise, namely LLMs. Because of that, the full-scale exercise has also been repeated in the pilot countries/languages.

The search for new skills terms follows the standard processing of an OJA, which includes lowercasing, removing stop words and punctuation and creating n-grams. N-grams allow multi-word skills to be analysed as a single term, for example 'problem-solving'. In the WIH DPS, skills are then matched to the ESCO classification. The algorithm presumes that new skills will be found in the text of OJAs close to existing ESCO skills terms, as they will be used for similar purposes (e.g. to describe a desired set of skills of a candidate).

Words (or terms) are represented by semantic vectors ('word embeddings'), usually drawn from a large corpus using co-occurrence statistics or neural network training. To ensure maximum coverage of possible new terms, both the preferred and alternative labels for ESCO skills will inform the search for candidates with new skills.

The steps in the new skills search are as follows.

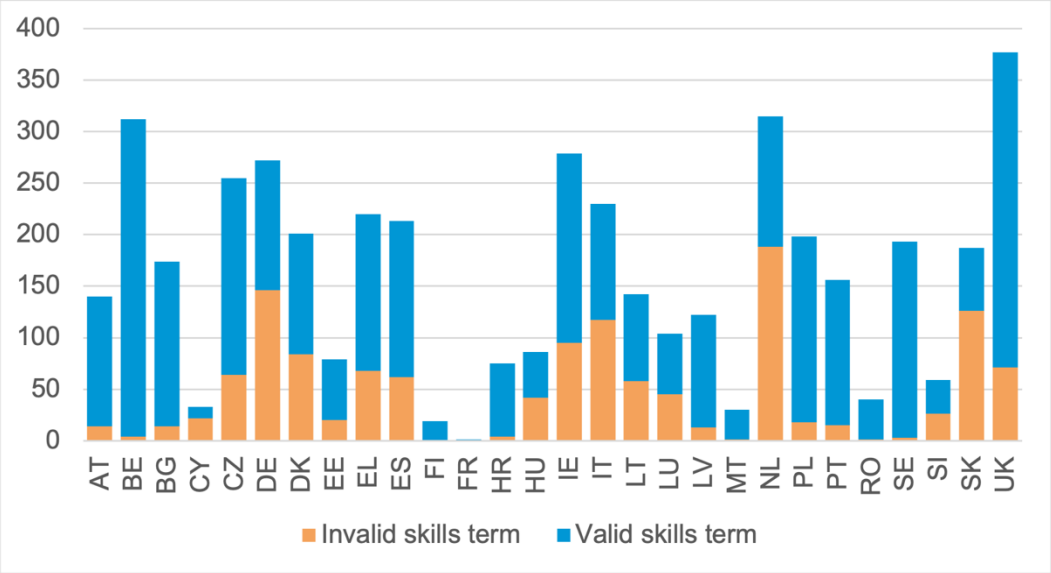
(a) Up to 10 terms that are most similar to an existing ESCO term were selected.
(b) Duplicates of preferred or alternative labels were dropped (since various and often synonymous terms will serve as the starting point, some duplicates of existing terms are inevitably labelled).
(c) The above point includes dropping terms deemed too close to existing skills, since they do not represent a novel term. This is done by considering morphological similarities between words. For example, the candidate term 'solve problems', which consists of letters that are very similar to 'problem-solving', will be discarded.
(d) All candidate terms that do not appear frequently enough were discarded. In our case, the threshold was set to at least one appearance in every 10 000 OJAs.

(e) As the word embeddings, along with the skills, may potentially also include considerable noise (manifesting as words that are semantically similar to ESCO skills but do not represent skills in the particular context), an LLM (in this case, the Mistral7b Instruct model) was deployed to identify whether the embedding represents noise or an actual skill in order to limit the otherwise substantial need for human term validation. The performance of the chosen LLM was boosted by a few prompt engineering techniques, especially:
    (i) defining the skill in the query;
    (ii) suggesting optimal responses (such as trying to limit them to yes–no responses to reduce the number of steps in the post-processing of the response of the LLM);
    (iii) providing examples of OJAs containing the skill as an example of the use of the term.
(f) The LLM-curated new skill candidate list was then uploaded to a human validation interface, which allows the language experts to:
    (i) select a country/language dataset;
    (ii) look at the LLM description of a candidate skill term and why it was considered a potentially new skill;
    (iii) look at five of the most similar ESCO skills per the word embedding model of the country's OJAs;
    (iv) make a suggestion for the type of skill (professional, transversal, digital).
(g) Based on the information provided, the experts then assessed the candidate terms, specifically:
    (i) whether they consider the term a novel skill rather than a term that is too generic or does not represent a skill;
    (ii) whether they agree with the type of skill proposed by the system (professional, transversal or digital);
    (iii) either selecting the ESCO skill from the five suggested options that are mostly similar and could be associated with the new candidate or deciding that none of the proposed skills is similar;
    (iv) selecting the type of the relationship the skills candidate has with the most similar ESCO term: novel, broad, narrow or alternative.

Overall, the system proposed some 4 500 candidate terms. Of these, the human experts confirmed some 3 200 (or 71%) as skills across all the countries. The most significant additions were made to the Belgian (most likely through the Dutch language pipeline) and UK datasets (over 300 new skill terms). The new skills structure pattern is valid for most of the countries, except France, Cyprus and

Finland, where the number of new skills additions was the lowest (1, 11 and 19, respectively) (Figure 10).

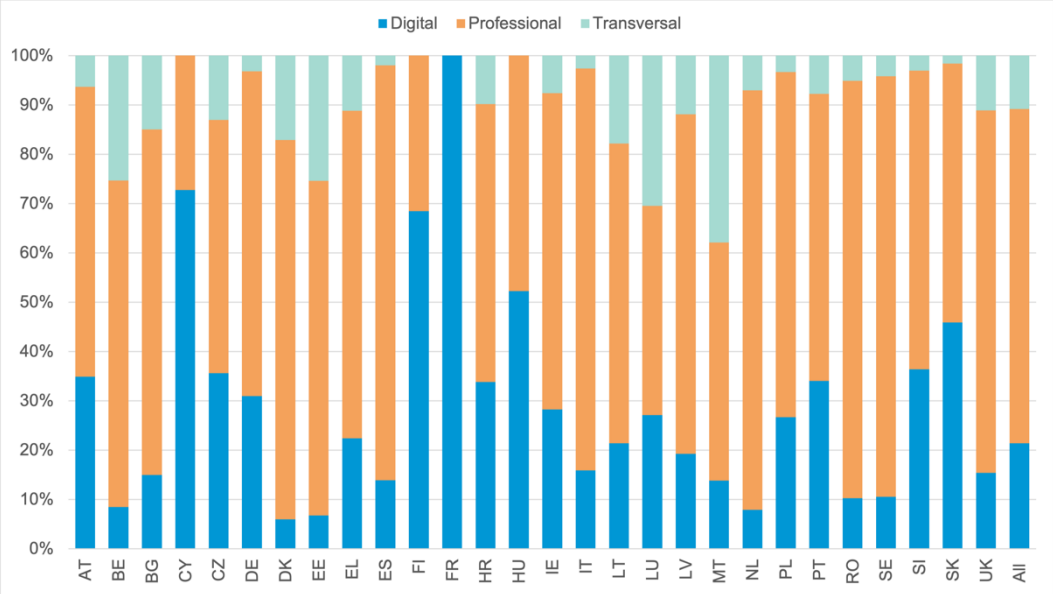
Figure 10. **Number and validity of new terms identified across countries**



Source: New and emerging skills recommendation system.

Most of the valid new skills terms come from the professional skills category (skills neither digital nor transversal) (Figure 11). There are considerably fewer new digital skills identified, which is understandable, given that these skills often appear in English, even in the OJAs written in other languages. The proportion of transversal skills is even less, but this is to be expected as they comprise the smallest group of skills in the ESCO classification.

Figure 11. **Types of valid new skills identified across countries**



All=all countries.

Source: New and emerging skills recommendation system.

An important part of the analysis of the proposed new terms was to assess the relationship between the new skills terms and ESCO skills terms. For half of these 3 200 new terms (1 645 terms) it was possible to establish such a relationship, meaning that they were classified as novel skills terms, broadening, narrowing or offering an alternative to an existing ESCO term. Most of the rest (41%, or 1 306 terms) were deemed to be 'not associated' with any ESCO terms. This means that, before these terms are included in the WIH DPS skills vocabulary, further assessment is needed to double-check and confirm their usefulness in broadening the scope of the ESCO classification. Finally, some 275 candidates (9%) could not have their relationship with ESCO skills terms assessed.

AI-driven ontology augmentation could significantly enhance the skills vocabulary utilised by the WIH DPS. Figure 6 highlights the uneven coverage of ESCO skills terms across national languages, with over 2 000 terms detected in the English language pipeline, while many southern and eastern European language pipelines operate with substantially lower numbers. However, the volume of newly detected skills terms (Figure 10) suggests that the new and emerging skills tool presented here can increase these figures by 20-30% for the most challenging languages, such as Bulgarian (estimated 32% increase), Greek (27%), Latvian (21%) and Czech (20%). Such an improvement would enable a more balanced skills analysis across countries and at the EU level, reducing the current

bias towards large countries with numerous OJAs and higher skills yields, such as Germany, the Netherlands, Italy and France.

The integration of LLMs into the WIH DPS tool portfolio has demonstrated its effectiveness in enhancing skills identification and matching. A pilot exercise conducted several years ago, when LLMs were still emerging, identified only tens of potentially interesting new terms per country, many of which were deemed invalid by expert assessment. In contrast, the current LLM-powered approach has yielded hundreds of new terms, with a success ratio of 71% (measured by the share of valid terms over all newly identified terms).

The skills augmentation activities outlined in this chapter do not require frequent repetition. These initiatives aim to address specific issues with skills detection in the WIH DPS and have achieved significant improvements. Although new skills emerge continuously, conducting skills augmentation activities every few years is likely to be sufficient, as the power of AI tools and methods is expected to increase substantially in that time, promising even better detection of valid terms.

4.5. Concluding remarks

Accurately identifying and categorising skills and occupations within OJA systems is critical for tracking labour market trends and responding to shifts in demand. While the ESCO framework offers a valuable foundation for building skills ontologies, its slow update limits its effectiveness in capturing the rapid evolution of the skills required in the modern workforce. Additionally, multilingualism introduces a significant challenge for cross-country skills extraction and comparison. Variations in language, terminology and cultural context make the creation of universally applicable skills categorisations challenging, complicating accurate data interpretation across countries.

To address these challenges, a dynamic blend of AI-driven automation and human-validated workflows is essential to ensure that our ontologies remain current and adapted to linguistic diversity. The approach can be even more fine-tuned when looking at specific groups of occupations or skills. This focused approach was successfully tested to identify skills and occupations that are key for the twin (green and digital) transitions (described later). Therefore, in subsequent phases of this project we may replicate the approach for the health and social care occupations or for STEAM – science, technology, engineering, arts and mathematics skills.

Furthermore, fostering a strong collaborative relationship with the ESCO team could help align their updates more closely with real-world trends and address multilingual considerations, ultimately enhancing the accuracy and responsiveness

of the system. This approach will strengthen the representation of skills in digital labour markets. Moreover, it will also support stakeholders – including employers, jobseekers and policymakers – in navigating and adapting to the constantly evolving and multilingual world of work.

Chapter 5.

Using online job advertisements to understand the digital transition

The digital revolution and its impact on the labour market and skills needs has recently been high on the EU policy agenda. However, as digital technologies change rapidly ([12]), keeping harmonised tools (such as ESCO) up to date is not easy. As the ESCO classification serves a purpose slightly different from that of providing an immediate picture of the skills needed in the labour market, it is natural that some skills will need to be reflected immediately in the taxonomy. For example, skills related to generative AI tools such as ChatGPT appear in ESCO v1.2, launched at the end of May 2024, despite ChatGPT being introduced in the autumn of 2022.

Keeping classifications such as ESCO up to date to ensure that they reflect current situations and emerging trends is a time-consuming and costly process that requires the launch of a public consultation and the involvement of labour market specialists or representatives of national institutions responsible for maintaining occupation and skills classifications in each Member State. For these reasons, the updates are not regular and do not keep up with the changes enough, especially in the digital transition, which is characterised by continuously and rapidly evolving technologies that translate into significant changes in the occupations and skills in demand.

Moreover, updating the DPS in line with ESCO v1.2 is also a time-consuming activity. However, the information extracted from OJAs indicates the demand for digital skills, and regular revisions of skills classifications will be needed to keep them up to date. The version of the ESCO classification used for the DPS ([13]) includes 1 201 skills and knowledge concepts that are labelled as digital. To address this challenge, we explored the potential of complementing the ESCO classification with information downloaded from two digital platforms. This chapter explains the details of the procedure followed, which includes applying LLMs to identify a set of terms related to the digital world and provides an adequate category of terms.

([12]) Every year globally more than 10 000 new applications for patents related to computer technology are submitted to the European Patent Office.

([13]) At the time of preparation of this publication (spring and summer 2024), the DPS was using ESCO v1.1.1.

5.1. Finding sources to keep classifications up to date

Finding a way to keep a classification up to date means finding a source of information that can provide both:

(a) updates and coverage of the domain (here, digital skills);

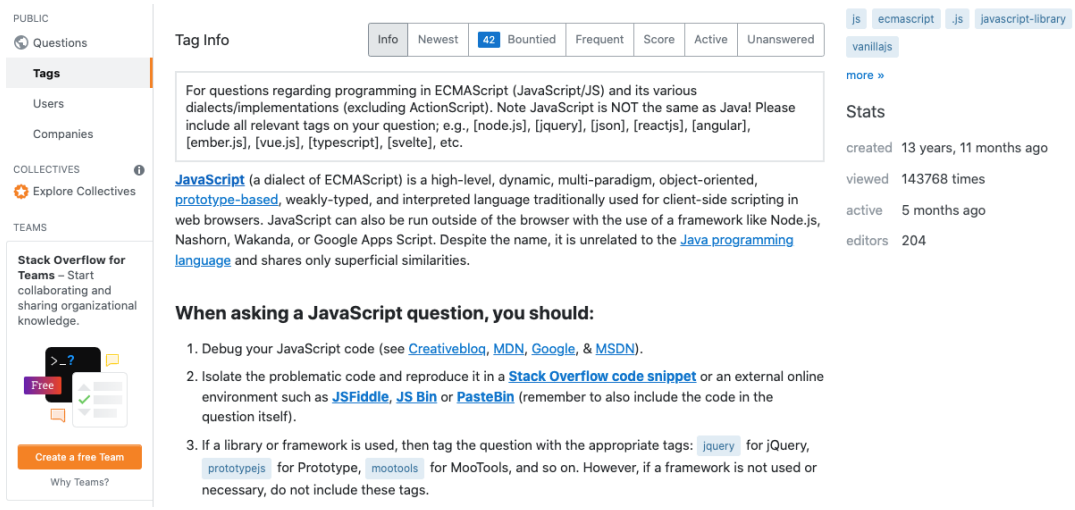
(b) information complete enough to allow the production of descriptions and statistics associated with each term that are similar to those already covered in the classification.

Two data sources were tested to provide regular updates for digital skills classification, namely Stack Overflow and GitHub. As the selection of these portals was made rather arbitrarily, the list could be extended to other specific portals (e.g. specialised portals for AI skills) in the future.

Stack Overflow is a community-based platform with many users ([14]). The main benefit of using Stack Overflow data for updating digital skills classifications is that all threads on the portal are flagged using unique tags. Tags are labels that categorise a participant's questions with other, similar, questions and make the navigation of this platform easier. For example, any question using the programming language JavaScript should be tagged '#JavaScript'. For each tag, a description is provided and validated by the community of users (Figure 12). The website includes information such as when the tag was created, which can be used as a proxy for when the technology was introduced. Furthermore, the Stack Overflow community of users associates each tag with a set of synonyms that can help identify some alternative labels and increase the efficiency of information extraction.

([14]) According to Stack Overflow growth and usage statistics (2024), as at November 2022 there were more than 100 million visitors to Stack Overflow every month.

Figure 12. **Information about the tag #Javascript provided on the Stack Overflow website**



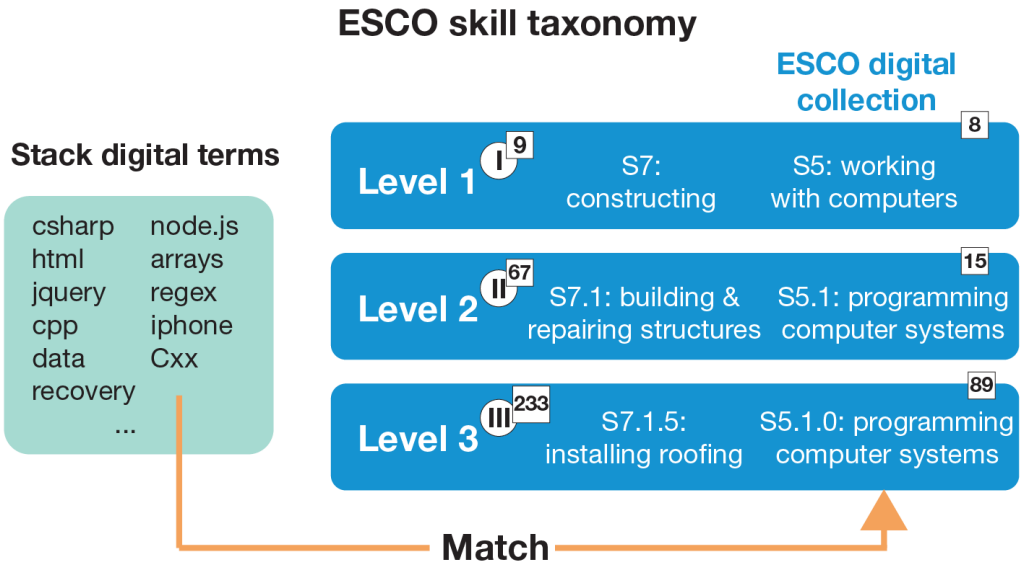
Source: https://stackoverflow.com/.

The GitHub portal users use a similar concept of tags to enable searching of repositories. The tags available from GitHub generally match those provided by Stack Overflow. Therefore, the information about tags extracted from GitHub was used to complement information obtained from Stack Overflow (e.g. description, release date) or to add missing information (e.g. about a number of public repositories matching the tag that can be used as a proxy for the level of adoption of the given technology). Combining the two sources provided approximately 65 000 different tags/technologies. The whole dataset was then processed to reduce the number of tags effectively duplicating very similar ones. This helped reduce the final number of tags from about 65 000 to around 40 000.

5.2. Exploring the suitability of new terms for updating classifications using large language models

The tags extracted from GitHub and Stack Overflow included 40 561 heterogeneous terms related to digital skills (e.g. JavaScript), digital knowledge concepts (e.g. machine learning, data mining), digital tools (e.g. Git), digital devices and others. To understand whether these tags are new relevant terms for updating classifications, they were first evaluated for their potential association with the list of ESCO digital skills and knowledge concepts that is a specific subset of the ESCO classification (we will call it the 'ESCO digital collection'). Only the 89 terms belonging to the three-digit level of the ESCO digital collection were considered Figure 13.

Figure 13. **Selecting terms from the ESCO digital collection for the matching procedure**

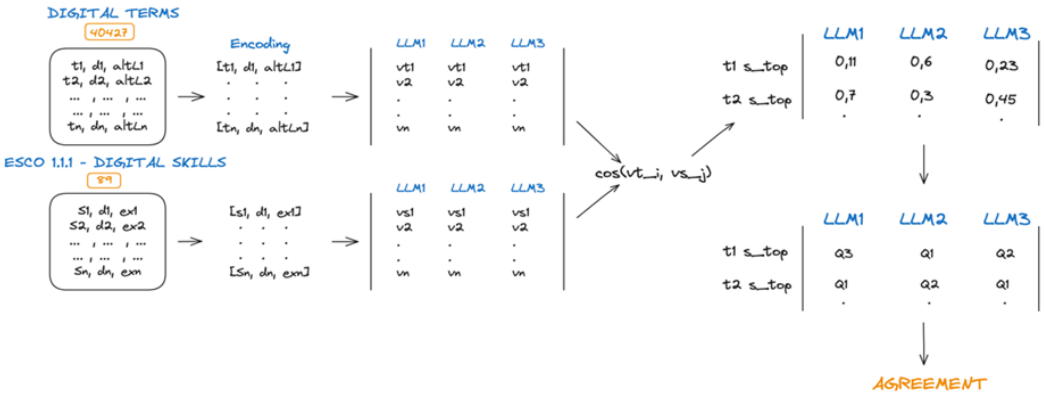


Source: Authors.

The approach used for matching the term involved the use of LLMs to obtain a vector representation of both digital terms and ESCO digital skills. A similarity measure was used to identify the most similar ESCO concept for each digital term.

In the absence of expert evaluation and to avoid relying on a single model, three LLMs were selected using the leaderboard available on Hugging Face. The first three free models were selected at the time of project development: bge-large-en-v1.5, ember-v1 and gte-large were used to minimise the risk of errors. The matching process is illustrated in Figure 14.

Figure 14. **Schema of the matching process using LLMs**



Source: Authors.

The LLMs undergo an initial encoding phase in which they transform the text provided by a human (in this case, the list of digital terms) into an internal representation (the so-called encoding phase). Eventually, each model's vector representation of each term is available, and the cosine similarity ([15]) can be computed to identify the highest score for similarity between each digital term and the ESCO digital skills. The results obtained were stored in a matrix containing the three values from each model that best matched each digital term. The matching results were compared to determine whether there was agreement on the best possible match between the models. The quartile method was used to identify a single match for each term ([16]).

In cases where all three models agreed on a unique match in the ESCO digital classification, that one was chosen as the match. When there was not a unique match, the following steps were applied:

(a) if a 'best matching term' was agreed upon within the quartiles, that term was the one taken;

(b) if there was not a unique best matching term and we had the same quartile for more than one match, then one was taken as 'preferred' and the other(s) as 'alternative(s)' (lines 2 and 3) (see Table 3).

([15]) Cosine similarity measures the similarity between two vectors of an inner product space. It determines whether two vectors are pointing in roughly the same direction.

([16]) Since the models work differently and obtain different vector spaces, we could not directly compare the scores given to each match. Therefore, the quartile distribution of the scores for each model was obtained, which allowed us to compare the results and determine the most suitable ESCO skill for each digital term.

Table 3. **Schema for selecting matches with multiple ESCO skills**

term_code	term	num_ matches	ESCO_skills	ESCO_ skill_code	preferred
STACKEXCHANGEv2 _100	File	2	[‘using word processing, publishing and presentation software - (Q1) - (1)’, ‘managing, gathering and storing digital data - (Q1) - (2)’]	['S5.6.2', 'S5.5.2]	Using word processing, publishing and presentation software
STACKEXCHANGEv2 _10001	Star-schema	3	[‘managing information - (Q2) - (1)’, ‘sorting materials or products - (Q2) - {1)’, ‘managing, gathering and storing digital data - (03) - (1)’]	['S2.3.0', S6.1.1', 'SS.5.2]	Managing informatlon
STACKEXCHANGEv2 _10002	Compass - geolocatlon	3	[‘operating communications equipment - (Q1) - (1)’, ‘browsing, searching and filtering digital data - (Q3) - (1)’, ‘monitoring environmental conditions - (Q3) - (1)’]	['S8.6.4', 'S5.5.1', 'S2.8.5']	Operating communication equipment

Source: Authors.

The construction of quartiles is beneficial in understanding the quality of matches produced by the models. For instance, 15% of all terms were mapped to a match that falls within the first quartile for all models. Approximately 40% of the terms were mapped with the highest possible similarity (falling in quartile 1 or 2) for all models. The application of this method resulted in a reduction in the number of potential new digital terms from 40 561 to a total of 28 687 digital terms, a possible enrichment of the ESCO classification.

5.3. Applying large language models to categorise new terms

The initial dataset consisted of various (not only digital) terms. Therefore, it was necessary to filter out terms that were irrelevant to the digital world. The list of 13 labels or categories taken from DBpedia – programming language, software framework, software tool, machine learning algorithm, computer programming keyword, operating system, digital device, computer hardware, computer network, database, file format, cloud computing and IoT – was used to enrich the ESCO classification of digital skills. Zero-shot classification, a machine learning approach in which a model is trained to classify data without having specific examples of all classes during the training phase, was applied. In other words, a zero-shot classification model can generalise and classify new data belonging to classes not

seen during the training phase ([17]). After running the model, a matrix that assigned a 'score' to each of the 13 category groups was obtained, indicating the probability of a term belonging to that class. To select the best labels for each term, we kept only those with a probability of at least 90% (see Table 7 in Annex 2). It is worth noting that 6 047 of the digital terms have no label above the selected threshold of 90%, which suggests that they may need to be more generic and could be excluded from the classification.

Categorising data has proven to be beneficial for various reasons. It allowed us to detect the presence of concrete machine learning libraries such as 'TensorFlow' but also general digital terms such as 'file'. It also allowed us to detect overly generic or misleading terms, which are not suitable for inclusion in our proposal for the updating of the digital skills taxonomy. For instance, terms like 'stopwatch' or 'smart TV' are classified as 'digital devices' by the model, but we may not want to include them in a digital taxonomy of skills.

5.4. Testing new terms for presence in online job advertisements

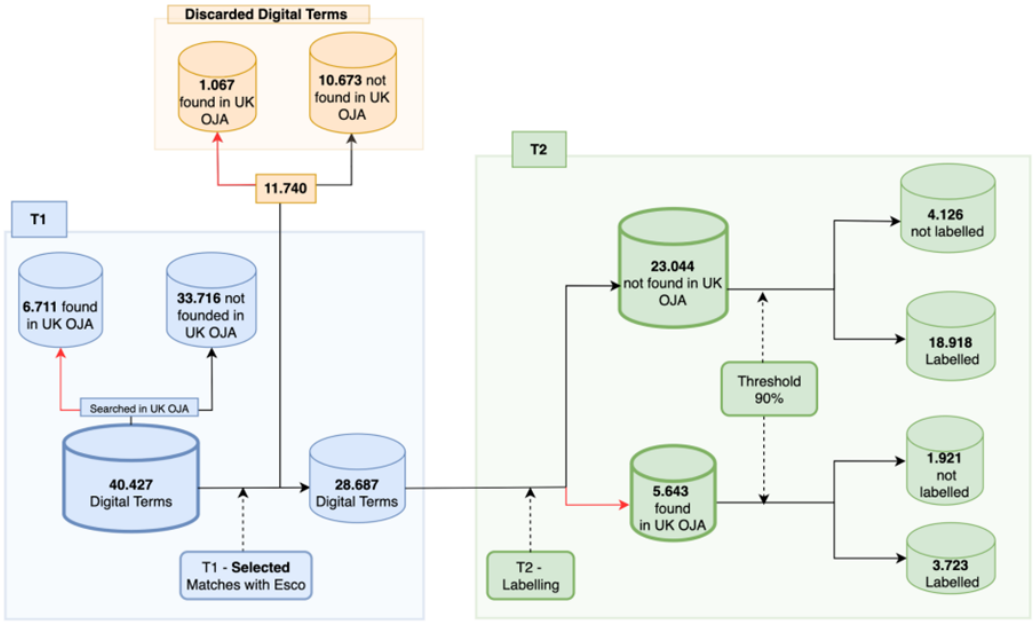
To further refine the choice of terms for the updating of the classification used in the DPS, additional information on which terms corresponded with OJAs recruiting for information and communications technology (ICT) roles was added ([18]). The diagram in Figure 15 illustrates the data flow. Starting with the original dataset containing 40 427 terms (excluding those already matched in the ESCO classification), only 6 711 had a match in the UK OJA considered for 2018–2023 (indicated by the red arrow in the diagram). During the procedure of matching terms with the ESCO digital collection using LLMs (T1 in the diagram), the number of terms was reduced to 28 687. Of this group, 5 643 had correspondence in advertisements published by employers in the United Kingdom. However, 1 067 terms with OJA correspondence were discarded from this first filter as belonging to something other than the digital world. Out of the terms selected after the procedure of matching terms with the ESCO digital collection that also have correspondence in OJAs in the United Kingdom, 1 921 were left out (not labelled), as the matching value with a digital label was below the selected threshold of 90%. This suggests that these terms may be too generic or are used in different contexts

([17]) The bart-large-mnli model, a version of BART trained specifically for the zero-shot classification task, was used in this task (https://huggingface.co/facebook/bart-large-mnli).

([18]) For a selected list of ICT roles, see Table 6 in Annex 2. OJAs from the United Kingdom (UK OJAs) were used as a benchmark dataset.

and have thus been excluded from the list of terms proposed for the classification update. Most terms are classified as 'software tools', 'computer networks' and 'programming languages'.

Figure 15. **Data flow, from the initial list of terms through the procedure of matching terms with the ESCO digital collection (T1) and finding the labels for the selected term (T2)**



Source: Authors.

5.5. Expert evaluation of terms

The final step in the selection of terms to be proposed for the updating of the ESCO classification includes human validation. Of the 5 643 terms with a match in the English OJAs, we focused on those assigned a single best label, for a total of 3 723 terms (Figure 15 above). Experts were asked to assess whether the term should be included in a digital skills classification. Specifically, they were asked to indicate whether it expresses a necessary or usable skill or proficiency in the ICT field. There were three possible answers to this question.

(a) A positive answer, indicating that the term can be proposed as valid and should be included in the classification, for example MIPS, an assembly language.

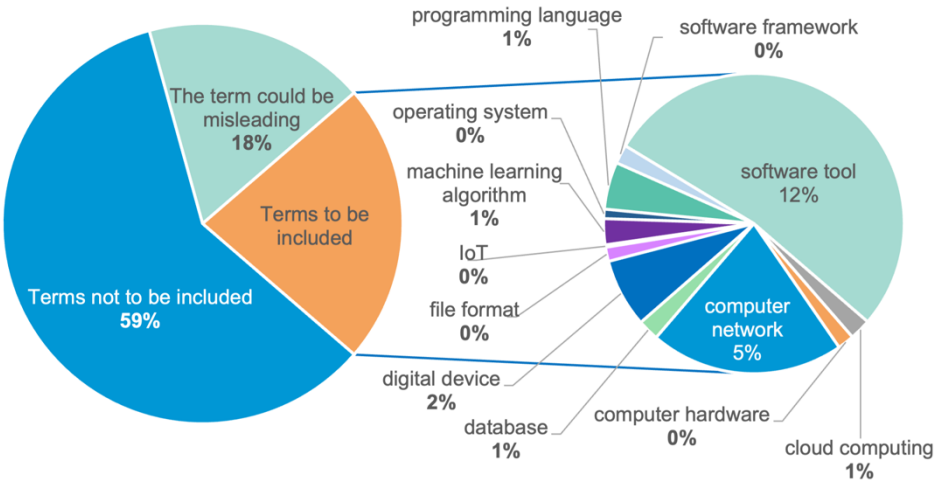
(b) A positive answer but one that points to possible challenges in information extraction (e.g. related to its ambiguity). For example, the term 'Pascal', which indicates a programming language, could also be used in the English language to refer to certain pieces of jewellery, or in physics as the SI unit of pressure;

another example is the word ‘class’, used in programming languages but also in many other contexts.

(c) A negative answer, indicating that the term should not be proposed for the updating of the classification; for example, JPEG and RGB are both terms referring to the digital world (a file format and colour values), but they do not represent skills or knowledge in this particular context.

The results of the human validation process are summarised in Figure 16. Despite a relatively high ‘discard rate’ (terms that were considered by experts as not suitable for inclusion or terms that could be misleading), we managed to acquire nearly 400 new terms that were accepted by experts as emerging digital skills that are valid for inclusion in our proposal for the updating of the classification. Half of those were software tools (e.g. TreeGrid, Checkmarx), 21% were related to computer networks (e.g. XMPP, RTCP) and 5% to programming languages (e.g. Haml).

Figure 16. **Terms by received category in the evaluation (left) and proposed new terms for relevant skills for updating the classification by the assigned category (right)**



IoT = internet of things.

Source: Authors.

5.6. Concluding remarks

In this chapter, we have described the steps in an approach applied to the collection of tags describing questions posted on the Stack Overflow platform and complemented with information obtained from GitHub, which allowed us to select the final list of terms that we can submit as our proposal for updating the ESCO

classification of digital skills. The application of LLMs proved useful not only for dividing terms into related and not related to the world of digital skills but also for finding a category for organising them (e.g. name of devices, name of keywords of programming languages and tools).

Chapter 6.

Using online job advertisements to understand the green transition

The urgency of addressing challenges related to the climate crisis has led Member States to agree on implementing the goals of the European Green Deal. This transition towards reducing greenhouse gas emissions is being achieved through a shift towards 'greener' models of production and consumption, including the implementation of a circular economy approach and investment in developing new technologies. The transition is changing the structure of the labour market and also creating new jobs for which appropriate skills (let us define them as 'green skills') are needed.

As the green transition is expected to happen over the next two decades, aiming to make the EU-27 climate neutral by 2050, and it will be accompanied by digital and demographic transformations, it is essential to understand the extent and nature of the expected changes in the skills required to better shape future education and training policies. With some restrictions, OJA data are one of the few most promising data sources for the identification of the occupations needed for a green economy (see Vona, 2021), and therefore they are also a good source of information in terms of the granularity of data and their timeliness for addressing various policy questions (see Cedefop, 2024). For example, identifying how similar the skills content of green and non-green jobs is can help determine the degree of up- or reskilling needed to enable the transition to the green economy (Bowen et al., 2018).

This activity was developed in late 2021 when the ESCO green skills and knowledge concepts classification was not publicly available. Therefore, the initial aim of this project was to produce a classification of green skills and related occupations based on the content of OJAs (data-driven approach). However, the ESCO classification of green skills became publicly available ([19]) during the process. This allowed us to test the terms extracted for semantic similarity with existing skills. The final list obtained through the application of AI techniques (particularly natural language processing and machine learning) contributed to the enrichment of the ESCO terms as either alternative labels or a proposal for a new skills term to be included in the forthcoming update of the ESCO classification.

([19]) The official release took place on 28 January 2022 (see Green skills and knowledge concepts: Labelling in the ESCO classification).

6.1. Data-driven approach to extracting green skills

Despite the growing importance of green skills for the economy, there is a lack of consensus on what type of occupations and skills should be considered green (see Auktor, 2021). One of the first systemised works on the classification of green occupations was developed under the occupational information network programme sponsored by the US Department of Labor / Employment and Training Administration. The green economy programme of O*NET, which started around 2009 (Dierdorff et al., 2009), was designed to collect detailed information on tasks for around 100 occupations more closely involved in the green economy and label them as 'green' or 'non-green'. The approach was driven by using academic journals, commissioned reports, industry white papers and governmental technical reports. In the O*NET occupational taxonomy, the 'greening' of occupations is defined through the expected impact on economic activities and technologies, which allows distinguishing between (i) existing occupations that are expected to be in high demand due to the greening of the economy – green increased demand; (ii) occupations that are expected to undergo significant changes in task content due to the greening of the economy – green-enhanced skills; and (iii) new occupations in the green economy – new and emerging green skills (ibidem).

The O*NET classification has recently been used to identify green jobs and skills via green tasks (Vona et al.; 2018; 2019 Vona, 2021). This has been done by exploiting granular O*NET data on the task content of occupations ([20]). In practice, there are several other approaches used to identify green skills (Consoli et al., 2016) depending on the analysts' focus on:

(a) process (e.g. green jobs will be related to waste management, waste treatment, energy use monitoring);
(b) products and services (e.g. skills needed to produce hybrid or electric cars, insulation products or energy monitoring systems);
(c) industry (e.g. manufacturing of energy-efficient appliances, filters or wind turbines);
(d) occupation (e.g. solar panel technicians).

([20]) It is important to highlight how the specific structure of O*NET allows this distinction. O*NET contains detailed information – coming from workplace surveys, occupational experts and expert analysis – on both tasks (e.g. what workers are expected to do at the workplace – the 'demand side') and skills (e.g. the abilities and competences that workers should possess to perform work tasks – the 'supply side') which receive an importance score on a scale of 1–5. In the O*NET database, each occupation, including the green ones, is defined as the combination of two vectors: a vector of scores in general skills (defined for all occupations by occupational analysts and experts) and a vector of dummies for the presence of text-rich, specific tasks (collected from job incumbent surveys for each occupation and then classified into green and non-green ones).

The uniqueness of the approach presented in this chapter is that it uses OJAs to define green occupations by their skills and not vice versa. In this approach, the skills shape an occupation and determine whether (and how much) it is green. Therefore, to define an occupation as green, it must contain green skills, and its greenness is determined by the presence of skills needed for a green economy.

6.2. Building a bottom-up data-driven approach

Following the bottom-up approach (see NESTA 2021), Cedefop has built its 'bag of green-related terms' by reviewing the following reports and documents: the Classification of Environmental Protection Activities (CEPA, 2000), the Classification of Resource Management Activities, the International Renewable Energy Agency Global Renewables Outlook 2020, LinkedIn, SkillsFuture Singapore, Joint Research Centre GreenComp, O*NET and ESCO green classifications. The initial list of 140 terms obtained from these publications was used to train a machine learning model to enhance it with lexicon variations as they emerge from OJAs.

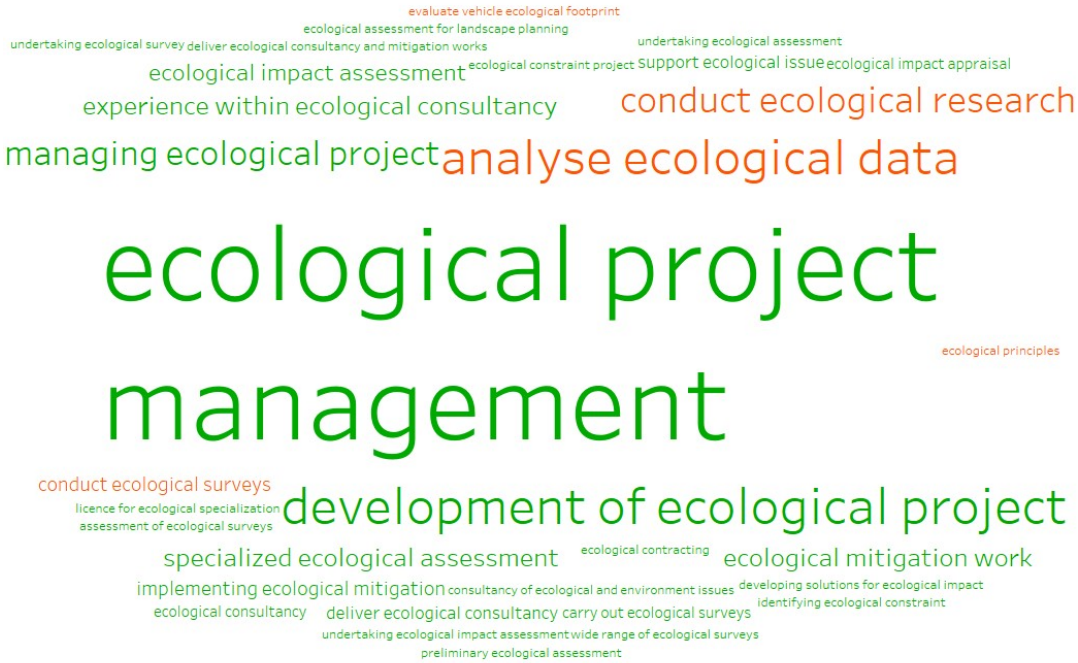
The computational approaches developed by CRISP, the Interuniversity Research Centre for Public Services of the University of Milano Bicocca (Giabelli et al., 2022) (i.e. a distributional semantics pipeline based on word embedding models), were used to build up a vector model ([21]) from 6 million OJAs collected in 2019 in the United Kingdom in the English language, using this initial list of 140 terms as a baseline for exploring OJAs for other mentions (words, terms, lexical jargon, synonyms, etc.) connected to green skills. When word embeddings are used, matches are made using the concept of cosine similarity between words. Similarity can be selected with different degrees of stringency. In this case, a threshold of 1 was adopted (the highest possible cosine similarity value); only exact matches were selected.

This deliberate selectivity was implemented to minimise the possibility of misleading results (due to false positives) and maintain the high precision of our taxonomy. For example, the term 'ecological' was linked to five different skills in

([21]) Several space vector models using various architectures were trained: Word2Vec, GloVe and FastText, state-of-the-art algorithms for generating word embeddings. Overall, we generated 260 models. Hyperparameter selection for each architecture was performed with a grid search over the following parameter sets: Word2Vec (80 models): algorithm ∈ {SG, CBOW} × HS ∈ {0, 1} × embedding size ∈ {5, 20, 50, 100, 300} × number of epochs ∈ {10, 25, 100, 200}; GloVe (20 models): embedding size ∈ {5, 20, 50, 100, 300} × number of epochs ∈ {10, 25, 100, 200}; FastText (160 models): algorithm ∈ {SG, CBOW} × embedding size ∈ {5,20, 50, 100, 300} × number of epochs ∈ {10, 25, 100, 200} × learning rate ∈ {0.01, 0.05, 0.1, 0.2}.

the ESCO taxonomy: ‘ecological principles’, ‘analyse ecological data’, ‘conduct ecological surveys’, ‘conduct ecological research’ and ‘evaluate vehicle ecological footprint’. This list was enhanced with 27 new proposals for skills terms (Figure 17). The list of green terms and a step-by-step description of the approach is available in Annex 3.

Figure 17. **Word cloud of ESCO taxonomy green skills terms that includes the word ‘ecological’ enhanced with terms identified in Cedefop bottom-up approach**



NB: The size of the word corresponds to its observed frequency in OJAs; orange indicates existing ESCO skills and green indicates terms found in OJAs with the proposed word embeddings approach.

Source: WIH-OJA data monitoring system.

For the term ‘sustainable’, which is linked to at least 19 different skills in the ESCO taxonomy, the method used resulted in only 10 new proposals of skills terms, for example developing innovative sustainable solutions or designing sustainable buildings (Figure 18).

Figure 18. **Word cloud of ESCO taxonomy green skills terms that include the word 'sustainable' enhanced with terms identified in Cedefop bottom-up approach**

NB: The size of the word corresponds to its observed frequency in OJAs; orange indicates existing ESCO skills and green indicates terms found in OJAs with the proposed word embeddings approach.

Source: WIH-OJA data monitoring system.

The final list of 182 English terms extracted (the United Kingdom was treated as a benchmark) was first translated by national experts for Germany, France, Italy and the Netherlands. In 2023, the project was scaled up to cover all languages in the WIH-OJA system. Lastly, using cosine similarity, each green skills term was associated with the ESCO taxonomy of green skills. In cases where such an association did not exist, the term was added as a new green skills term.

6.3. Human input assisting green terms extraction

Collaborating with country experts played a pivotal role in enhancing the quality of the work on green skills, offering valuable insights into the strengths and areas for improvement in the approach. Firstly, the role of experts was to validate the translations of terms from English to national languages. Specifically, the experts were asked to validate the 'mentions' as they emerged from OJAs and their translation into English and to refine/modify the translation if needed. Finally, experts were also used to evaluate the coherence of the green skills found in the

OJAs by examining their alignment with the occupations with which they were associated in the advertisements.

Using a Likert scale, where 1 means ‘strongly disagree’ and 5 ‘strongly agree’, experts were asked if they found any inconsistencies in the green skills associated with occupations and if they could identify which of the skills were problematic. Using this approach, experts identified some green skills/terms that are missing (e.g. ecologists). As all ontology-based extractions become quickly outdated and do not include the newest terms, the detection of some skills that, for example, are related to emerging green technologies could be omitted. This, however, is expected to be of negligible magnitude, as such skills will eventually be detected by new emerging skills algorithms (see Section 4.4.2).

6.4. Concluding remarks

This work confirmed possible differences between top-down and bottom-up approaches to skills extraction. The overall expert observation was that some ‘green by definition’ occupations lack green skills terms. It may be the case that, in advertisements for such green occupations, the green element is already included in the job title. For example, for environmental engineers or environmental protection professionals, employers may skip mentioning other green skills in the description of the requirements for the position. The challenge of applying this bottom-up approach to the identification of green occupations is related to the fact that very often green terms not linked to skills requirements are present in the part of the advertisement that describes the company mission or vision, for example ‘as a company, we’re committed to driving energy efficiency and addressing the global emission challenge’ or ‘the company’s mission is to create a sustainable brand that increases the environmental awareness in our society’.

This may lead to issues in the extraction of skills terms. To avoid this bias, an additional step was introduced to help identify green terms related only to the job requirements and not used to describe the company’s mission or vision. Each OJA was converted into a set of texts that were more likely to encode skills to reduce the noise and improve the overall embedding quality. The idea was to use sentinel words as anchors for a window that should contain only the skill-related part of the OJAs to restrict word-embedding processing to sentences about skills to remove the bias. Overall, as we followed a very conservative approach based on exact matching, all our findings should be considered lower bound, providing a robust foundation for analysing green jobs in the online job market.

Chapter 7.

Extracting information about the field of study

While the concept of 'skills-based' hiring has been promoted recently to address skills shortages ([22]), qualifications still play a significant role in employers' recruitment decisions. Together with using them for monitoring the impact of the digital and green transitions on jobs and skills, qualifications are an essential topic of Cedefop's research and analysis.

A qualification is understood as 'a formal outcome (certificate, diploma or title) of an assessment process certifying that an individual has achieved learning outcomes to given standards and/or possesses the necessary competence to do a job in a specific area of work' (Cedefop, 2008). The qualifications provide information, ranging from the level of education, which may also demonstrate knowledge of a particular field and mastery of relevant skills, to candidates' commitment to professional development and alignment with the company's objectives.

In the hiring process, qualifications are used to screen candidates and filter out applicants who do not meet the minimum requirements for the position. They also reduce hiring risks related to potential staff turnover and decreased productivity. They are often an indicator of a prospective employee's suitability for a role, as employers can assess candidates based on their qualifications. Moreover, qualifications also work as quality assurance because they can provide a standardised way of measuring candidates' knowledge, skills and competences. Finally, they help ensure compliance with legal or regulatory requirements in certain industries or for specific roles (see, for example, NCVER, 2005).

In job advertisements, qualifications are usually expressed as a combination of level of education and field of study. Information about the levels of education is already extracted from the OJAs and mapped to the eight groups of the ISCED (see ISCED, 2011). To better understand qualification requirements in OJAs, Cedefop conducted an exploratory exercise aimed at mapping how employers specify the field of study.

This chapter explores two approaches – an ontology-based extraction and a data-driven method – to obtain more detailed information about the required field of study. Understanding the academic or professional domains individuals

([22]) See 'Addressing talent shortage with skill-based hiring' or 'Skills-based hiring: Tackling the labour shortages'.

specialise in is crucial for various applications, including educational planning, workforce development, talent management strategies and research analysis.

7.1. Ontology-based extraction of the field of study

All WIH-OJA classification pipelines (for skills, occupations, sectors, etc.) are based on ontology-based matching models. The development of a new pipeline to extract information about the required field of study started with the search for existing classifications that would be suitable for this purpose. The International Standard Classification of Education – Fields of education and training (ISCED-F) framework appeared promising, as it is used for various purposes, including providing educational statistics and collecting data, making international comparisons of education and training programmes and facilitating the recognition of qualifications and degrees across countries.

ISCED-F is an international framework developed by the United Nations Educational, Scientific and Cultural Organization (UNESCO) to categorise and classify fields of education and training at various levels. ISCED-F is part of the ISCED ([23]) framework, which aims to provide a globally recognised and standardised way of classifying education systems and programmes. ISCED-F specifically focuses on categorising fields of study and training programmes and is organised into three hierarchical levels, each representing a different degree of detail in classifying fields of education.

For example, the broad group categorises fields of study into 11 groups (e.g. natural sciences, social sciences, humanities, engineering and technology). The broad group includes more detailed narrow subcategories (around 29 labels). For example, within the broad field of natural sciences, there are narrow fields like biology, chemistry and physics. Detailed fields within the narrow field of biology include microbiology, genetics and ecology. There are approximately 80 labels for detailed fields in the ISCED-F classification.

ISCED-F is the basis for developing the knowledge pillar of the ESCO classification; in this way, it is also covered by the WIH-OJA. The knowledge pillar, however, contains over 7 000 terms ([24]) – many times more than the detailed labels of the ISCED-F classification – and a single OJA may contain many of them.

([23]) ISCED is the reference classification for organising education programmes and related qualifications by levels and fields of education.

([24]) See European Skills, Competences, Qualifications and Occupations (ESCO): knowledge.

The challenge of the OJA classification is to identify which of the many possible terms is likely to indicate the field of study specified (Figure 19).

Figure 19. **Knowledge and skill terms mapping in an OJA**

Qualifications/skills/attributes

- Degree in biomedical engineering, mechanical engineering or similar discipline with 3+ years medical device industry experience in a product development or manufacturing role
- Experience with embedded C and C++ and embedded fundamentals.
- Experience with Microcontrollers and popular interfaces such as Bluetooth, UART, etc.
- Proficient in using Solid Works, Altium or equivalent 3D CAD design package.
- PC literate (word processing, spreadsheets, database) and a good knowledge of project management tools.
- A good understanding of quality management system principles (e.g. ISO 9001 or ISO 13485) and/or FDA quality system regulation processes according to 21 CFR part 820.
- A team player with excellent interpersonal and communication skills, with the ability to solve problems ad hoc.

NB: Orange indicates knowledge terms, and blue indicates skills terms.
Source: Authors.

The main difficulty of using ISCED-F for information extraction is the fact that it is not formally adopted across all languages of the WIH-OJA database. The ISCED-F provided by UNESCO is available in English, French and Spanish. For some other European countries, the official national fields of study classifications are available via their NSIs. To build the field of study classification across all languages covered by the system, we used the official language versions, either from UNESCO or from the NSIs, and we added non-official lists of fields of study in several Member States that were identified via web searches. Finally, all non-English pipelines have been boosted by automated translations from English (Table 4). Multiple tools for translating the English anchoring terms into other languages were used to ensure the quality of the translations ([25]). Outputs from individual tools were compared, and the final translation was based on a 'majority vote'.

The second challenge in using ISCED-F for the field of study extraction is related to the uneven number of available terms across languages. While the Spanish list, built on the UNESCO and the NSI lists, contains over a thousand terms, the Czech, Dutch, French, German, Hungarian and Swedish lists have only

([25]) DeepL, Google Translate, Reverso Context, M2M100 (Hugging Face model), Grammarly.

around three hundred terms (Table 4). Yet, most language versions, for which no official national list has been identified and have thus had to rely on translation from the English list, contain only 153 terms. The results of the extraction using this ontology-based classification method are compared with the data-driven method described in Section 7.2 (Figure 20).

Table 4. **Availability of field of study classifications and the number of unique terms across languages covered by the WIH-OJA system**

Language	Official (ISCED)	Official (NSI)	Other	Automatic translation	Number of unique terms
bg				X	153
cs		X		X	275
cy				X	153
da				X	153
de		X	X	X	305
el			X	X	153
en	X	X	X		187
es	X	X		X	1 036
fi				X	153
fr	X	X	X	X	268
hr				X	153
hu		X		X	282
is				X	153
lt			X	X	153
lt				X	153
lv				X	153
mt				X	153
nl		X (from Statbel)		X	276
no				X	153
pl		X		X	219
pt			X	X	153
ro				X	153
ru	X		X	X	153
sk				X	153
sl				X	153
sv		X		X	262

Statbel = Belgian statistical office.

NB: Please see Annex 4, Table 9, for detailed information about sources.
Source: Authors.

7.2. Data-driven approach to the field of study

Acknowledging the insufficient coverage of the ISCED-F classification for extracting information about the field of study, similar to extracting skills from the ESCO classification, Cedefop decided to develop a data-driven approach. This approach started with building the list of 'anchoring terms', defined as terms likely to appear close to preferred fields of study qualification requirements in the content of an OJA. For example, in the first sentence in Figure 20, the term 'degree' is an anchoring term, as it introduces the employer's requirement for a candidate to have a degree in biomedical engineering, mechanical engineering or a similar discipline.

Figure 20. **Examples of anchoring terms (in green) found in various OJAs**

- **Degree** in biomedical engineering, mechanical engineering or similar discipline.
- An electrical national **craft certificate** is essential.
- A **completed upper secondary education** in construction or engineering.
- A **decree** 50/78 col. & 5 is mandatory.
- CORY **registration**.
- **Chartered** accountant.
- **Diploma** in hotel management.
- Six Sigma **Green Belt**.
- **Certified** energy manager.
- Google Data Analytics **Professional Certificate**.

Source: Authors.

This list of anchoring terms was first developed in English by in-house Cedefop experts based on the review of OJAs, and synonyms were searched for in OJAs written in other languages. To ensure that the terms found in OJAs are correct with reference to the corresponding English term, the list was validated by Cedefop ReferNet experts ([26]). The list includes the following terms: 'degree', 'certificate', 'knowledge in', 'qualification', 'professional experience', 'license', 'education', 'vocational education', 'craftsmanship', 'technical understanding' and 'trained' but also terms such as 'required', 'appreciated', 'preferable', 'necessary', 'suitable' and other similar ones.

The list of anchoring terms was used to extract possible field of study terms that appear in their proximity. The circle of proximity was defined as three words preceding and five words following the anchor, excluding stop words. Over 21 000

([26]) More information on ReferNet is available online.

combinations of anchor and possible field of study terms were retrieved from a sample of the OJAs in all languages. These combinations were passed to Cedefop ReferNet experts for their evaluation. The experts were requested to say whether the combination is a valid one, that is, whether it refers to the field of study or not. It was also possible to choose the option 'unsure' if the information provided in their decision was insufficient to make such a judgement (Table 5 gives some examples).

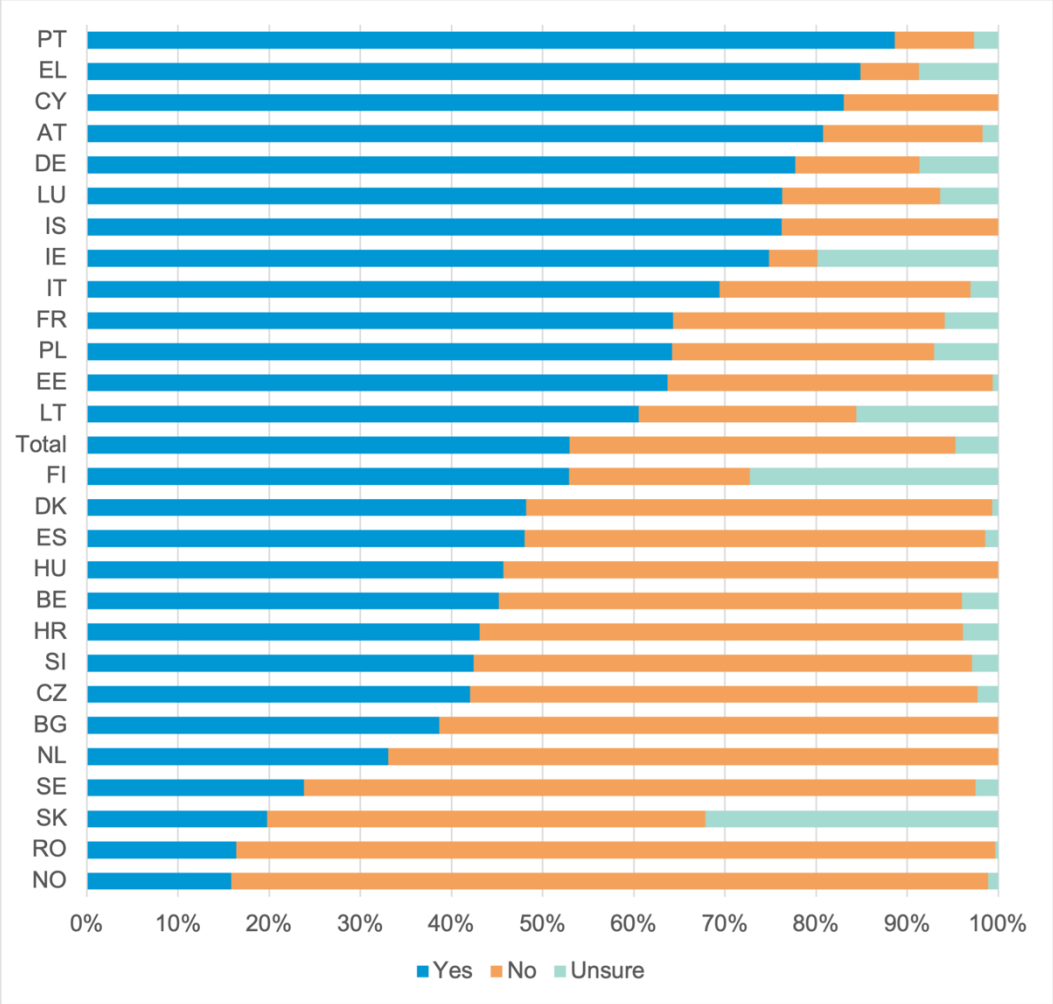
Table 5. **Examples of the outcomes of ReferNet's Irish expert's validation**

Outcome	Example
Valid: contains the field of study requirement	'qualification' and 'social counselling'
Invalid: does not contain the field of study requirement	'craftmanship' and 'sports'
Unsure: the information provided is insufficient to make any judgement	'medicine' and 'license'

Source: Authors.

Overall, the share of valid field of study term combinations reviewed reached 53%. The highest share was observed in expert reports from Portugal, Greece and Cyprus, while the lowest values were reported by experts from Norway, Romania, Slovakia and Sweden (Figure 21).

Figure 21. **Results of the ReferNet experts' validation of terms**



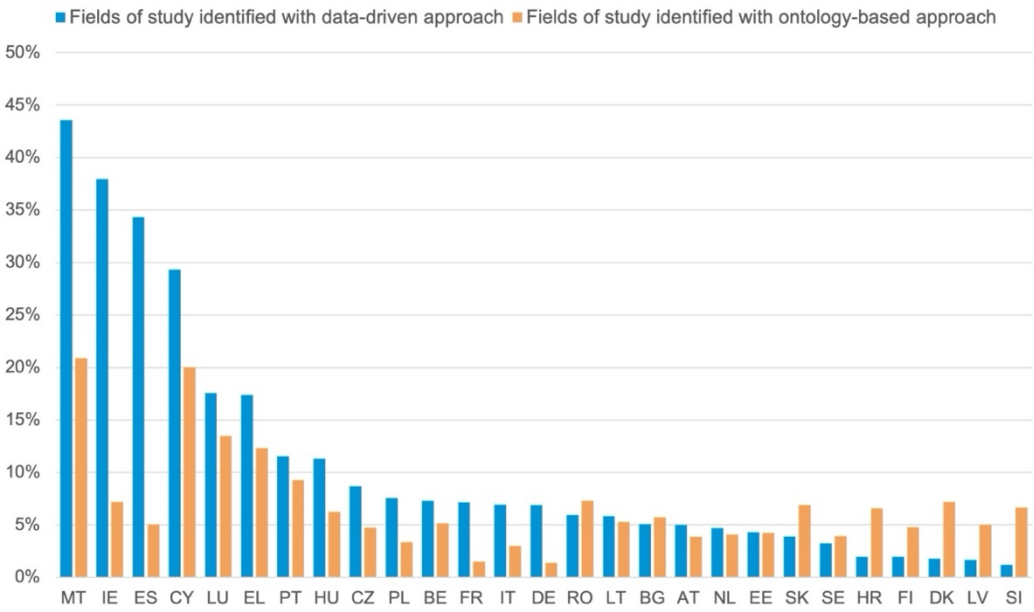
Source: Authors.

The outcome of the experts' validation (a fine-tuned list of terms) was used in the next step to extract the final information on the field of study from OJAs. Combining national language terms and anchors in the extraction process could be a more comprehensive approach. This approach could capture the various ways that fields of study may be mentioned in job postings, improving the accuracy of their extraction. It is essential to note that the validation process was conducted using the combinations of anchoring terms and fields of study based only on the OJA sample.

This validation method provided a fine-tuned list of terms that improved the accuracy and reliability of the final data extraction process for some countries like Malta, Ireland, Spain and Cyprus (Figure 22). However, it is essential to acknowledge that, if certain combinations were missing from the validation set,

they were also omitted from the extraction process. Therefore, for countries like Slovenia, Latvia, Denmark and Finland, the samples in terms of number of observations were too small to allow the data-driven approach to obtain similar levels of information extraction to those observed when the ontology-based classification approach was applied.

Figure 22. **Shares of OJAs with the field of study present by method of information extraction applied: ontology-based approach with ISCED-F classification vs data-driven approach with anchoring terms**



Source: Authors.

7.3. Concluding remarks

Further refinement of the data-driven approach presented in order to obtain information about the field of study from the content of OJAs is required before the approach can be implemented in the information extraction pipelines. Creating a training dataset will be necessary to address the challenges of differences in the fields of study extraction process across countries caused by small sample sizes. A training dataset created by experts who would annotate the raw content of OJAs, marking information about the field of study and anchoring terms, would improve the precision of the classifications.

Chapter 8.

Developing skills intelligence from online job advertisements

Cedefop defines skills intelligence as 'the outcome of an expert-driven process of identifying, analysing, synthesising, and presenting quantitative and/or qualitative skills and labour market information. These may be drawn from multiple sources and adjusted to the needs of different users' ([27]). As the content of OJAs captures the latest trends in occupations in demand and skills required and reflects employers' preferences, they offer a valuable source of information that can be used for building skills intelligence.

Although converting OJA data into skills intelligence is a complex task relying on advances in big data analytics, machine learning and natural language processing methods, the information extracted from the body of OJAs has already been used to contribute to various types of labour market analysis.

(a) Occupational analysis. For example, analysis of the occupation of computer scientist (Grüger & Schneider, 2019), various types of analyst positions (Nasir et al., 2020) and public health jobs (Watts et al., 2019).
(b) Sectoral analysis. For example, analysis of the information technology (IT) sector (Ternikov & Aleksandrova, 2020), tourism (Marrero-Rodríguez et al., 2020) and manufacturing (Leigh et al., 2020).
(c) Regional skills intelligence. For example, the skills requested in OJAs were compared with the skills taught in training and education programmes in Umbria, Italy (OECD, 2023).
(d) Skills in demand. For example, to understand which professions are demanding AI or green skills and to forecast the demand for skills (de Macedo et al., 2022).
(e) Time series analysis of trends in skill set requirements. For example, analysis of trends in entrepreneurial skills (Prüfer & Prüfer, 2019), foreign language skills (Fabo et al., 2017) and transversal skills (Pater et al., 2019) and of changes in skills requirements for journalists (Dawson, et al., 2021).
(f) To understand the impact of the green transition (Saussay et al., 2022).
(g) To measure competition among employers (Ascheri et al., 2022).

([27]) See, for example, key terms in the terminology of European education and training policy.

Moreover, OJAs can shed light on changing patterns in labour demand and so contribute to our understanding of shortages within specific sectors and occupations. This can be used by educators to tailor curricula to better align with the evolving needs of the economy, equipping learners with the competences and knowledge relevant for future employment opportunities. In addition, OJAs can play a crucial role in updating classifications of skills and occupations. By analysing job postings across different sectors and industries, researchers can identify emerging job titles, skills clusters and occupational trends that may necessitate revisions or expansions of existing classification systems.

Lastly, OJAs can be leveraged for short-term forecasting of skills trends (e.g. attempts to enhance the Cedefop STAS system – Short-term anticipation of skills trends and VET demand), providing timely insights into immediate skills demands and emerging job opportunities in the economy. By monitoring fluctuations in job postings, employers' hiring preferences and skills requirements, stakeholders can anticipate shifts in labour market dynamics and proactively respond to changing workforce needs (Cedefop, 2024). Overall, skills intelligence, which insightfully combines traditional data with OJA data, can provide valuable information for stakeholders, such as policymakers, educators, employers and individuals, allowing them to make informed decisions about skills development, education, training and career planning.

In this chapter, we will show a few ways in which analysing the content of OJA data can contribute to a better understanding of various dimensions of the impact of the twin transitions reflected in the changing demand for skills and occupations.

8.1. Better understanding of what occupations are in demand

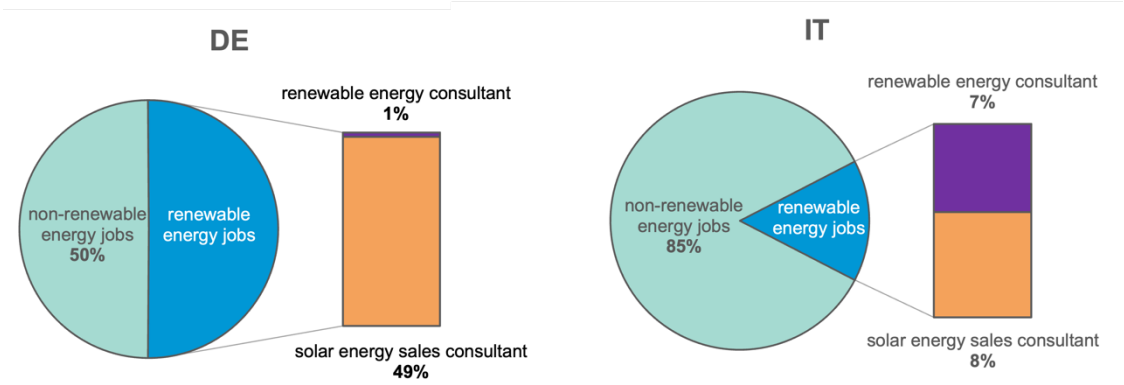
The granularity of information is the main advantage of using the content of OJAs in analysis of changes in occupations in demand when compared with the standard sources of information (e.g. surveys or administrative data). Although not every job title contains enough information to be classified according to ISCO, for many advertisements, a match with a corresponding occupational group can be found. In the WIH-OJA system, the machine learning-based classifiers are applied to find the best match between the job title and the existing lowest four-digit level of ISCO.

Even information at the four-digit level of ISCO will not capture many contemporary job roles in energy, as the last ISCO update took place in 2008. Jobs for renewable and solar energy sales consultants, which are currently classified as technical, like medical sales professionals, are cases in point. Looking at the five-digit level, there is enormous variation between countries when it comes to the

composition of this group. In recent years, jobs for renewable sales professionals' accounted for half of the OJAs for technical and medical sales professionals in Germany; in Italy the proportion was only 15%.

Nevertheless, even information at the four-digit level of ISCO may not capture the full extent of many contemporary jobs when considering such niche occupations as green jobs. For example, ISCO 2433, technical and medical sales professionals, contains ESCO job titles, such as ISCO 2433.3, renewable energy consultant, and ISCO 2433.5, solar energy consultant. Analysing data from OJAs can allow us to identify occupations at this level of detail in cross-country comparisons. Combining the data with information on the energy sector, we can observe that half of the advertisements for renewable energy consultants or solar energy consultants in Germany and 85% in Italy were not linked to the renewable energy sector (Figure 23).

Figure 23. **Structure of demand for technical and medical sales professionals by five-digit occupations in Germany and Italy**



Source: WIH-OJA data monitoring system.

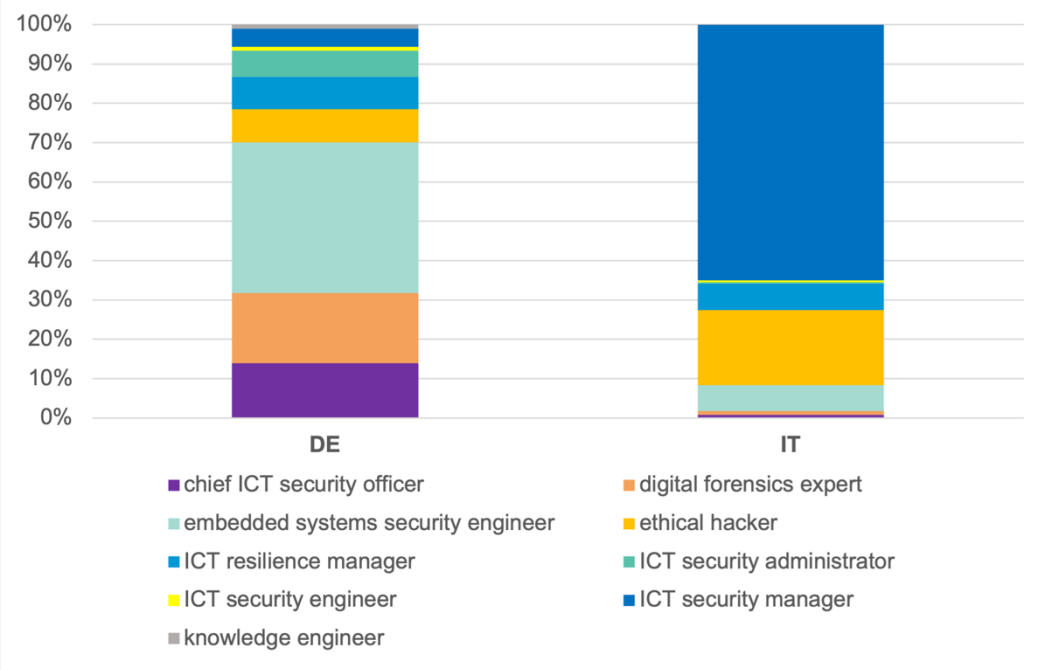
The analysis of OJAs driven by ontologies (such as those based on ISCO) depends on the updating of the underlying classifications. Therefore, many roles pertinent to the transformation of the economy due to the twin (green and digital) transitions have yet to be included in the classification, as they appeared in the labour market after it was already established (the last update of ISCO was in 2008). Consequently, these roles are currently categorised as 'occupations not elsewhere classified'. For example, an alternative fuels engineer who deals with liquefied natural gas, liquefied petroleum gas, biodiesel, bio-alcohol, hydrogen and fuels produced from biomass is currently classified as 'engineering professional not elsewhere classified'. Between 2018 and the end of 2022, this 'not elsewhere classified' group accounted for 12% of the total number of green occupations, amounting to 420 000 advertisements in 2022 alone (Cedefop, 2024).

Similarly, many IT roles, for example software tester or ICT test analyst, belong to the group 'software and applications developers and analysts not elsewhere classified' (ISCO 2519); similarly, ICT security administrator or ICT security engineer are included in the group of occupations 'database and network professionals not elsewhere classified' (ISCO 2529). As IT occupations have been in the labour market for longer than emerging green occupations, the share of the 'not elsewhere classified' group is considered to be 4% of all advertisements targeting IT workers. Still, as the market for IT professionals is much larger than that for green occupations, this 4% means that, on average, there are around 175 000 openings with unknown occupational profiles. This poses a challenge for analysing the demand for occupations crucial to the transforming economy. Until ISCO is updated again, the only way to obtain more detailed information about occupations that have emerged in the meantime and are becoming crucial for the twin transitions is to search for alternative classifications of occupations or to use the ESCO classification at deeper than the four-digit level ([28]).

For example, the outcome of classifying occupations at the five-digit level of the ESCO classification for 'database and network professionals not elsewhere classified' shows that in Germany embedded systems security engineers were most searched for by employers, while in Italy the most searched for occupation was ICT security managers (Figure 24).

([28]) The ESCO classification uses the same structure as ISCO up to fourth level of the hierarchy and adds more granular disaggregation at the lower levels.

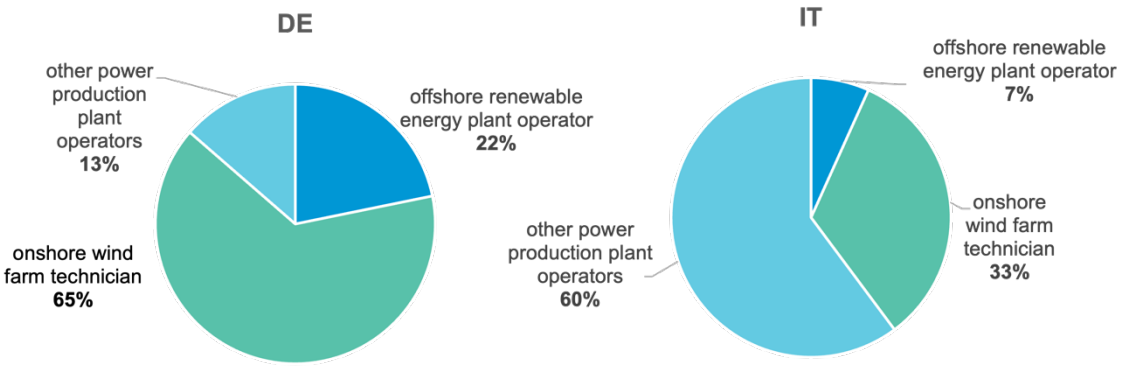
Figure 24. **Structure of demand for database and network professionals not elsewhere classified in Germany and Italy**



Source: WIH-OJA data monitoring system.

The exercises described in Chapters 5 and 6 could allow us to classify OJAs according to the ESCO (five-digit level) classification and thus allow better insight into occupations in demand, but sometimes even that level of granularity does not suffice. For example, the power plant operators (ISCO 3131) group can be split into three groups at the five-digit level: onshore wind farm technicians, offshore renewable energy plant operators and other power plant operators (including geothermal power plant operators, solar panel plant operators, hydroelectric plant operators). As the third group includes a variety of power plant operators, accessing five-digit level occupation data would allow us to better understand the structure of demand in Germany. In fact, only 13% of demand belonged to the group of other power plant operators. However, this share was much higher at 60% in Italy (Figure 25). Therefore, in the case of Italy, either the job titles need to be classified at a lower, six-digit, level of ESCO classification or the analysis would need to link occupations with skills to allow a better understanding of the percentage of employers recruiting solar panel, hydroelectric and maybe nuclear plant operators.

Figure 25. **Structure of demand for power plant operators by five-digit occupation codes in Germany and Italy**

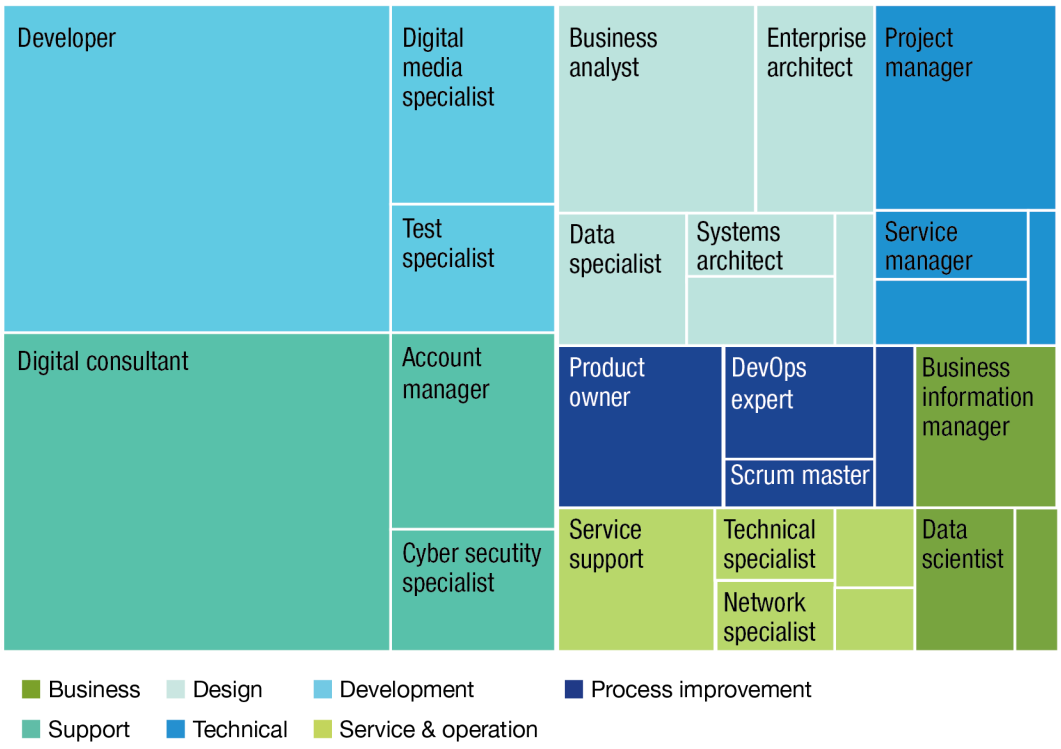


Source: WIH-OJA data monitoring system.

Another approach could be to use an alternative classification system. For example, in the case of the IT sector, the application of the European Committee for Standardisation (CEN) and the European Committee for Electrotechnical Standardisation European ICT Profile Family Tree classification, which includes 30 European ICT professional role profiles proved to be beneficial compared with the ESCO five-digit classification. The full implementation of detailed (five-digit) occupations in the DPS will be finished in the first half of 2025. The CEN-based classifier for OJAs written in English was used in a feasibility study (Figure 26). This approach allowed us to obtain information about 30 different IT occupations and their grouping into seven distinct ICT family profiles (e.g. business, design, support).

These families were created based on existing skills shared among IT professions, making this categorisation very useful for enterprises, human resources departments and professionals engaged in skills development. For example, training providers might use this more detailed information about skills families in demand to better tailor their education offers.

Figure 26. **Structure of demand for ICT workers in OJAs published in English by ICT family profile**



Source: WIH-OJA data monitoring system.

8.2. Understanding skills requirements in emerging occupations

The rapid pace of technological change is reshaping industries and transforming the nature of work. This evolution creates new jobs while reducing the number of others and requires workers to adapt and acquire new skills. This translates into the need for skills intelligence to address the questions related to emerging occupations and their skills requirements. For instance, as blockchain technology is more widely adopted across various industries and services, a range of blockchain-related occupations are appearing in the jobs market. The demand for blockchain technology-related skills, which is still a niche skill set in the labour market, was investigated using online job postings as a key resource (Chaise, 2021). In addition to proficient blockchain developers capable of constructing and managing blockchain networks and applications, non-technical roles such as business process developer, value chain architect and blockchain consultant have been identified (Chaise, 2021). Cedefop's analysis of over 3 000 OJAs published

in the EU identified two more relevant occupations: blockchain data analyst and blockchain chief security officer. These occupations, among others, required knowledge of blockchain technology that employers expected to be used in developing new solutions for them.

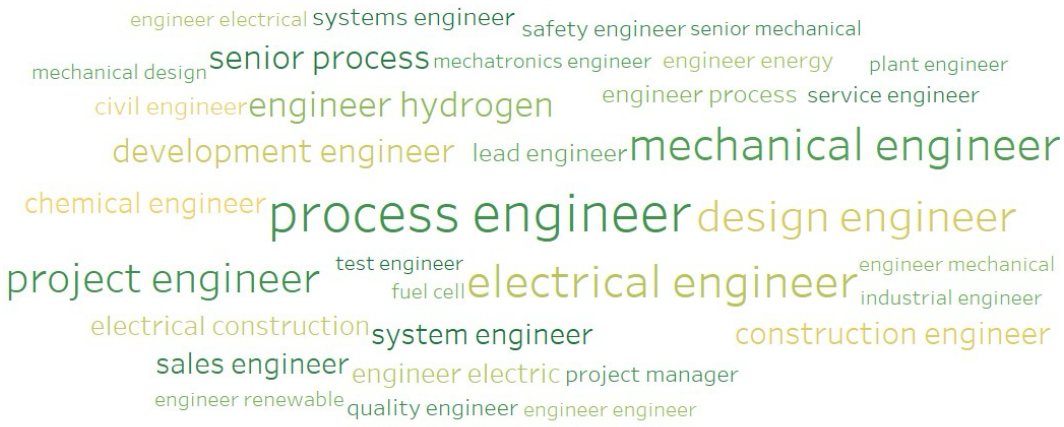
The content of OJAs can also be used to explore the emerging green occupations expected to play a significant role in the transition to a sustainable economy. OJAs may help identify new and evolving roles in areas such as green finance and sustainable urban planning or occupations related to new green technologies. For example, investing in the hydrogen sector holds significant potential for economic growth and job creation. Projections suggest that, by 2030, the hydrogen industry could generate around 1 million jobs in Europe alone ([29]). However, the transition to an economy based on hydrogen will happen only if adequate training and upskilling are provided for people.

Therefore, investment in skills should go hand in hand with investments in the hydrogen sector. Interviews with stakeholders confirmed that the hydrogen sector offers a diverse range of occupations with at least 200 distinct job roles required within the hydrogen value chain ([30]). The analysis of OJAs recruiting for the hydrogen sector can help us understand what types of occupations are in demand and what skills they require (Figure 27). For example, an analysis of the most common job titles from advertisements recruiting for roles related to hydrogen production shows that most of them target various engineering professionals. The roles most frequently targeted were process engineers, who are part of the technical team that brings cutting-edge products to life. Using OJAs we could also identify occupations less frequently targeted but still crucial for the hydrogen sector, for example mechatronics engineers. Among other responsibilities, these professionals are involved in evaluating performance and production quality and diagnosing and addressing production and process issues in the hydrogen sector.

([29]) See *Hydrogen roadmap Europe – A sustainable pathway for the European energy transition*.

([30]) See the *European Hydrogen Skills Strategy*.

Figure 27. **Word cloud showing the most common bigrams (two consecutive words) in job titles from OJAs recruiting for roles in the production of hydrogen**



NB: The size of the word corresponds to its observed frequency in OJAs.
Source: WIH-OJA data monitoring system.

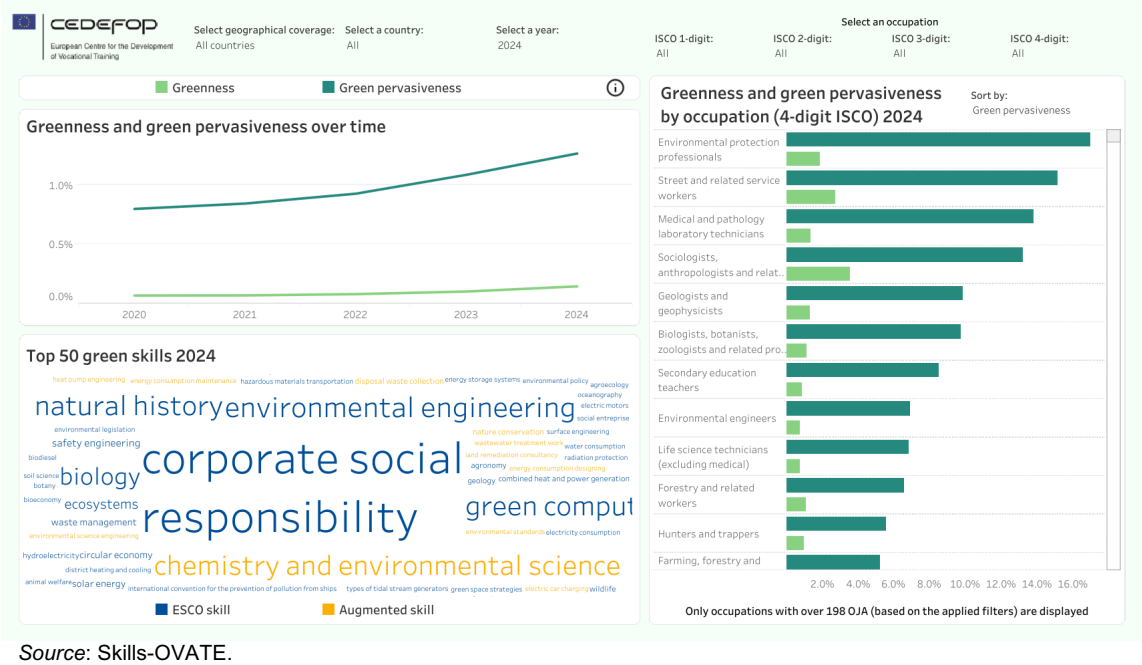
8.3. Understanding employers' changing needs: indicators of skills demand

Although there are many composite indicators and scoreboards available on the European Commission website that allow monitoring of the progress of the European Green Deal, only a few include a variable that allows the importance of human capital's involvement in this green transition process to be measured. The content of OJAs, with some restrictions, is considered one of the promising sources of information about the changing demand for green skills in the EU (Vona, 2021).

To address this gap in the monitoring of the green transition's impact on the labour market and to better understand how each occupation is impacted by the 'green transition' using OJAs, we constructed two indicators: green pervasiveness and greenness. Green pervasiveness measures the prevalence of green skills in the OJAs. It is calculated as the ratio of all OJAs with at least one green skill requested by employers to the total number of OJAs in the category analysed (e.g. at the occupational, sectoral and country levels). The greenness indicator compares the number of green skills requested by employers with the overall number of skills found in the advertisements for the occupation. The first indicator allows us to understand the percentage of OJAs that demand green skills and the second to understand for which occupations green skills are most important. Both indicators can help to assess the progress and impact of green policy initiatives on the labour market.

For example, Figure 28 shows the growing trend in green pervasiveness and greenness across all countries covered by DPS. The figure also presents 'the greenest' occupations measured by green pervasiveness (the share of OJAs requiring at least one green skill in the total OJAs for that occupation group).

Figure 28. **Greenness and green pervasiveness in OJAs**



Source: Skills-OVATE.

In a similar way to greenness and green pervasiveness, we can construct indicators that help us assess the impact of the digital transition on the skills required by employers. For example, digitalness ([31]) could be used to measure how important digital skills are for a single occupation by looking at the ratio of digital skills to the overall number of other skills required, and digital pervasiveness could also be considered to calculate the percentage of OJAs that demand at least one digital skill. In the case of digital skills, we may also be interested in understanding how digitalisation translates into the demand for various levels of skills.

For example, the digital skills level can be divided into three groups (high/medium/low) according to the level of digital skills required based on the classification developed for Cedefop's second European skills and jobs survey (ESJS2). In this classification, low-level digital skills include internet browsing, use of email and social media, writing or editing text and use of spreadsheets at work. Medium-level digital skills include skills in using specialised software, preparing

([31]) The same concept as for greenness was also applied to digital skills.

presentations and advanced use of spreadsheets. High-level digital skills include managing or merging databases, programming and coding skills and the design of IT systems and hard- and software (see example in Cedefop, 2023a, p. 61).

Another way of building an indicator of digital skills is to use the ESCO classification of digital skills and their linkages with occupation profiles to group them according to the share of digital skills they require (digital intensity). This enables us to distinguish occupations with low digital intensity, in which the share of digital skills required is between 1% and 3%, from occupations with high digital intensity, in which the share of digital skills is more than 5% (see example in Cedefop, 2023a, p. 60).

The content of OJAs was used in many analyses to understand the increase in demand for AI-related skills. For example, Back et al. (2021) evaluated whether OJAs can be used to monitor technology adoption, confirming their usefulness in understanding the variation in AI-related skills demand over time and sectors. Acemoglou et al. (2022) used vacancies seeking workers with AI skills as a proxy to understand trends in the uptake of AI activities across establishments with various levels of exposure to AI. They showed that higher exposure to AI affects the types of skills demanded by establishments, significantly reducing demand for some of the skills previously sought and increasing the emergence of new skills in employers' requirements.

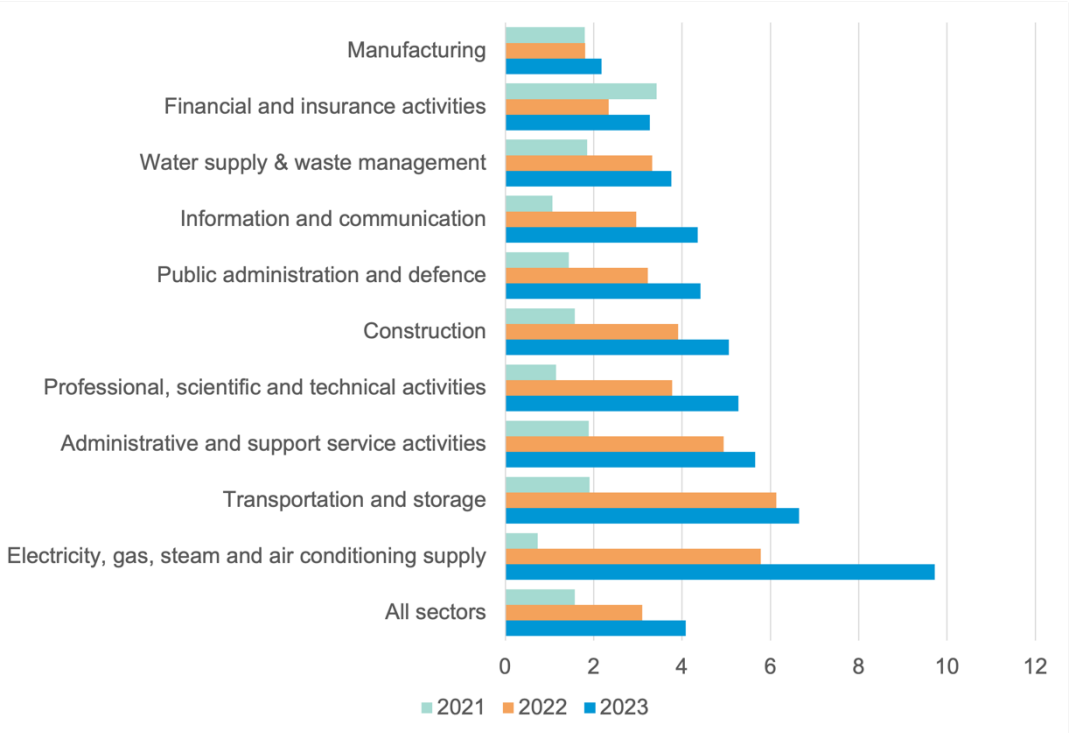
Cedefop analysis shows that the demand for AI skills has spread across a large set of occupations and sectors (Cedefop, 2023b). The difficulty with this type of analysis is that the list of terms for AI-related skills constantly changes, with many new terms emerging each year. Therefore, for any digital skills analysis, applying an ontology-based classification may lead to underestimating the demand (see Napierala, 2024), and bottom-up solutions might be more efficient in extracting information (see Chapter 5).

8.4. Tracking the demand at the sectoral level

The content of OJAs may also serve as a valuable tool for monitoring the changes in skills at the sectoral level. This can be showcased, for example, in the analysis of the speed of adoption and range of circular economy principles adopted by employers across Europe. The mentions of circular economy in the content of OJAs may indicate that the company advertising a vacancy has implemented this approach into their business or production processes. Beyond the term 'circular economy', other terms such as 'life cycle analysis' or 'industrial symbiosis' may signal a demand for workers because of companies' departure from mainstream 'linear' production processes.

The analysis of the sectors for which these advertisements were recruiting may be a proxy for the demand for skills for the circular economy. For example, Figure 29 demonstrates that, between 2000 and 2023, most sectors reported growth in OJAs, mentioning the circular economy term, but the highest growth was observed for the sector employing workers responsible for supplying electricity. This might reflect the recent accelerated transition to renewable energy sources, such as wind, solar and hydroelectric power, in most Member States.

Figure 29. **Growth in the prevalence of the term 'circular economy' in OJAs by sector, 2021-23 (base year 2020)**



NB: Only the top 10 sectors are listed. The X axis represents the frequency of mentions of 'circular economy' terms.

Source: WIH-OJA data monitoring system.

8.5. Detecting emerging changes in skill sets in demand

The analysis of skills mentioned in OJAs can also be used to detect the impact of introducing green or other policies on selected occupations and the skills they require. Developing a measure of the change in the skill set of an occupation is not easy. The skill set, in fact, can change because some skills become more or less important and new skills become important for an occupation. For example, as businesses and organisations are motivated by various policies (e.g. the circular

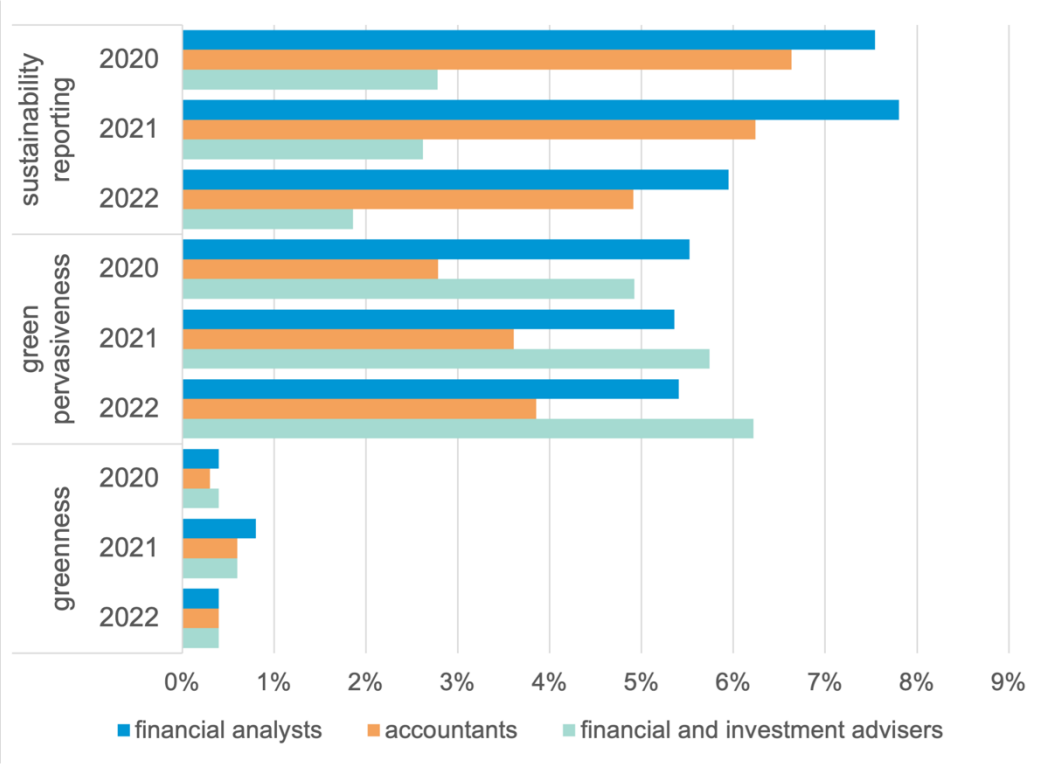
economy action plan) to integrate sustainability principles into their operations, regardless of their primary focus, we may observe some changes in the skills requirements for these roles.

An industrial designer (ESCO 2163.1), who, according to the ESCO classification, is responsible for the development of new ideas and their implementation into designs and concepts for a wide variety of manufactured products and does not require any green skills, is a good example. In Germany, between 2019 and 2022, on average, 4% of advertisements recruiting industrial designers requested at least one green skill, with environmental engineering being the most requested. As the skills of industrial design engineering combine systemic thinking with the creative process, this greening of formerly non-green occupations may indicate that German employers are shifting their interest towards more green ways of production. Industrial design engineers are professionals who specialise in applying scientific and technical knowledge to create products, materials or services in a creative way that is environmentally friendly.

Another example of using OJAs to detect the greening of non-green occupations is related to implementing the corporate sustainability reporting directive. Under this new directive, which came into effect on 5 January 2023, the inclusion of specific sustainability information alongside financial results became a mandatory requirement for companies. This regulatory change is anticipated to have an impact on finance and accounting occupations, necessitating adjustments in job holders' skill sets to meet the new requirements.

During the period analysed (Figure 30), a slight increase in the prevalence of green skills among accountants and financial advisers was observed. However, the average requirement for green skills in finance professionals' roles remained relatively low – below 6%. Despite the marginal prevalence of green skills in finance roles (greenness), which sits below 1%, sustainability reporting experience already ranks among the top 10 green skills employers seek in finance professionals' roles.

Figure 30. **Demand for green skills in finance professionals' roles in the EU-27, 2020-22**



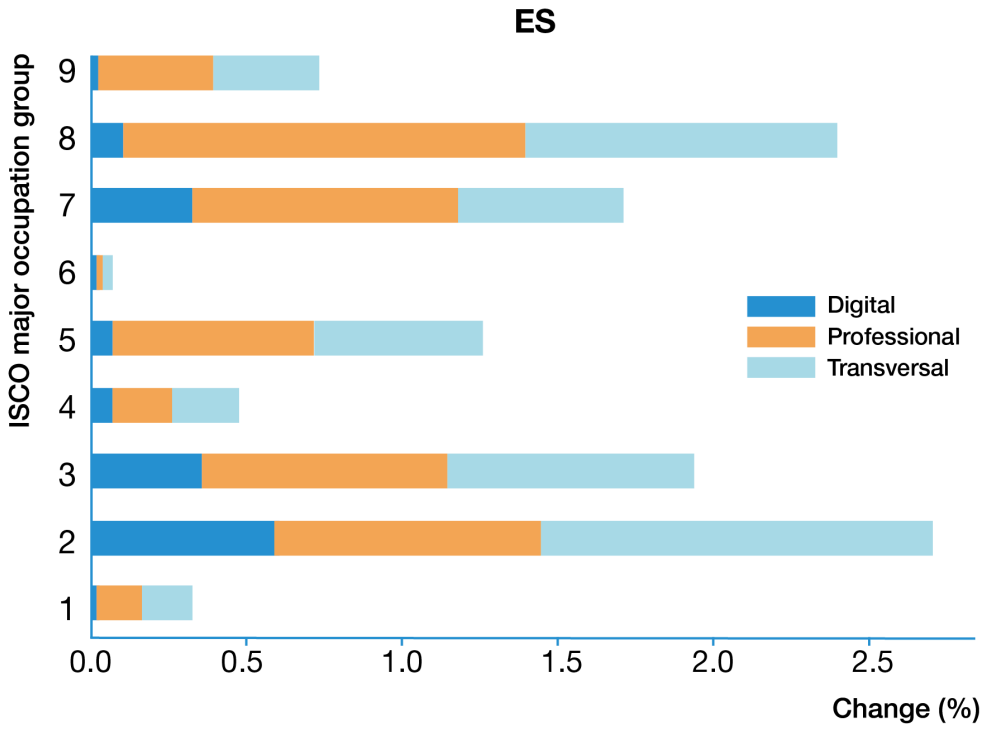
Source: WIH-OJA data monitoring system.

The examples described above represent a simple approach to understanding how skills requirements are adopted over time in response to changes in the policies implemented, based on measuring changes in the shares of OJAs containing a specific skill. Compared with this simple approach, the normalised revealed comparative advantage (NRCA) approach allows us to describe the developments in skills trends in occupations over time in a more refined way.

The change in skills specialisation within an occupation or group of occupations can be broken down into two dimensions: (i) intensity (or volume) of change and (ii) determinants of change, or which skills (or types of skills) influence the changes in the demand for skills in that occupation the most ([32]). For example, in Spain, between 2019 and 2022, the skills demand changed most for professionals (ISCO 2) and machine operators and assemblers (ISCO 8) (Figure 31). In the case of professionals, transversal skills drove the change the most, while professional (technical) skills were behind the change in the skills specialisation of machine operators and assemblers.

([32]) The normalised revealed comparative advantage (NRCA) index measures the degree of deviation of an occupation's skills specialisation from its expected value, scaled by the overall degree of specialisation.

Figure 31. **Intensity of change in skills specialisation and type of skills changed by ISCO (one-digit) occupations in Spain, 2019-22**



Source: Authors.

8.6. Gaining insights into skills in demand through certificate analysis

The content of OJAs is a valuable source of information that allows insights into occupations, skills required and trends detected at the sectoral levels. It also includes information about the desired levels of education or certificates the candidate needs to be equipped with. While information about fields of study is not yet extracted as a regular variable in data production (Chapter 7), an analysis of the certificates required for cybersecurity roles shows the potential for OJAs to be used to gain more understanding about the skills required (Figure 32).

A certificate is a credential awarded to individuals who have demonstrated proficiency in and knowledge of various aspects. For example, in the case of cybersecurity occupations, the possession of relevant certification is sometimes required with a combination of verifiable university coursework or a number of years of work experience. These certificates prove the holder's skills in network security, ethical hacking, incident response, digital forensics, risk management and compliance. Cybersecurity certification programmes can be broken down into two

main categories: professional cybersecurity certification programmes and academic cybersecurity certification programmes.

The former is designed for people already working in cybersecurity who need to be trained on some of the latest tools and software to detect, prevent and combat cybersecurity issues (e.g. CompTIA Security+). The latter are designed to give students the necessary skills and experience to get started in the growing cybersecurity industry. In 2022, the Certified Information Systems Security Professional (CISSP) ranked among the most sought-after certificates in the EU. Holders of this certificate, granted by ISC2, which is the world's leading members association for cybersecurity professionals, have demonstrable experience in IT security and capability for designing, implementing and monitoring a cybersecurity programme. The US National Institute of Standards and Technology cybersecurity framework certificate and security information and event management (SIEM) certificate were those most often required for cybersecurity roles after CISSP. Both tools help organisations to understand better, recognise and address potential security threats and vulnerabilities and improve their management of cybersecurity risks.

Figure 32. **Word cloud showing the intensity of the certificates mentioned by employers recruiting for cybersecurity roles in the EU in 2022**

NB: NIST = US National Institute of Standards and Technology; SIEM = security information and event management tool. The size of the word corresponds to its observed frequency in OJAs.

Source: WIH-OJA data monitoring system.

8.7. Linking online job advertisement data with other sources

Skills intelligence, by definition, may be drawn from multiple sources. In the literature, there are already a few examples of successful linking of information

obtained from OJAs with other sources of information. For example, Babina et al. (2023) leveraged a unique combination of two datasets. Using the company name, they merged information derived from CV data that captures the stock and characteristics of current employees with information obtained from OJAs that allowed them to estimate the demand for new employees and for their skills. The merging of these two sources of information allowed the authors to better understand how variations in workforce skills composition influences company investment in AI. Their analysis also sheds light on the impact of AI on the labour market.

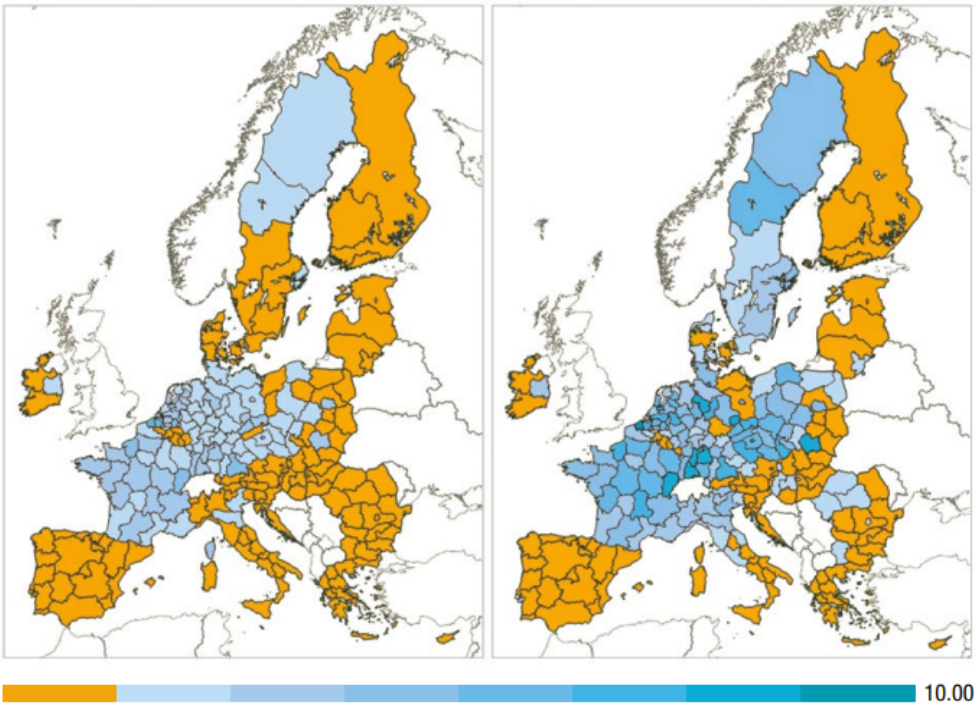
Similarly, Bennett et al. (2022) also used merged information from OJAs and CVs, with the only difference being that the information was obtained from one job platform in Uruguay, and the CVs were obtained from job applicants using this platform. The authors intended to evaluate if merging of these two sources can address the lack of longitudinal data on skills to study skills dynamics. They concluded that such joint datasets entail granular and longitudinal information, allowing meaningful analysis of both labour demand and supply at the skills level.

There are also examples of studies matching job postings with firm-level data. For example, Chen and Li (2023) merged job postings with information from a sample of US public firms for which firm-level historical characteristics, including financial information, were obtained from external databases (Compustat and CRSP). This merging of information allowed the authors to show that recruitment intensity, measured by time spent on recruiting, is higher for public companies expecting higher profitability. The companies prolong recruitment campaigns for high-skilled jobs to ensure that they reach high-quality workers.

Although company names are collected in the WIH-OJA data system, they are anonymised for statistical confidentiality purposes and, for that reason, are not available for carrying out similar analyses. Still, to get richer information about the current situation in the labour market, the information from OJAs can be used with information from other sources, for example the Labour Force Survey. As an example, combining information about the characteristics of the unemployed population with information about the skills required by employers at the regional level may allow a better understanding of skills mismatches (Figure 33).

The indicator for labour market tightness, calculated as the number of vacancies available (e.g. tertiary educated workers) divided by the number of unemployed (e.g. people having the required tertiary level of education), can indicate whether the market is tight (values above 1), in which case we observe more job offers than available candidates who could fill in the position, or not tight (values below 1), in which case we have more candidates than available posts.

Figure 33. **Overall labour market tightness (left) and tightness for tertiary educated workers (right) in EU regions, Q4 2022**



NB: The tightness for Q4 2022 was calculated based on the regional numbers of people unemployed in 2021. The scale is from 0 to10, with values below 1 indicated in orange.

Source: Cedefop (2023a).

8.8. Concluding remarks

As shown in this chapter, skills intelligence based on the content extracted from OJAs may contribute to various types of relevant labour market analysis, and therefore it may in many ways improve our understanding of the processes in the labour market that shape the changes in the skills requested of workers. Firstly, more granular information allows us to better understand the demand in the job market in terms of more detailed information about occupations in demand, but also about their skills (e.g. proxied for by certificates requested). Secondly, OJAs can contribute to a better understanding of the dynamics of the labour market, including regional variations in skills demand. Thirdly, information from OJAs, when merged with information about the supply side of the market, can provide insights into the disparities between the skills demanded by employers and those available in the workforce, helping to identify areas for skills development or training.

Chapter 9.
Conclusions and next steps

This publication wraps up 10 years of effort invested in establishing a multilingual and modular DPS for analysing OJAs under the leadership of Cedefop and Eurostat. Over the past decade, the DPS has evolved into a sophisticated system, integrating various data collection methods – such as web scraping, crawling and via API access – each adapted to the specific needs of diverse job advertisement sources. Through careful data preprocessing, which includes cleaning, deduplication and relevance filtering, the system minimises noise and redundancy, ensuring that only high-quality data are analysed. These refined data are then processed using language-centric ontologies and machine learning models, providing nuanced insights into job market trends across multiple languages.

Quality assurance is a cornerstone of the DPS, underscoring a rigorous approach to data reliability and accuracy. Regular validation practices, supported by expert review, help refine ontologies, classifiers and dictionaries in iterative cycles. By establishing evaluation datasets and setting a gold standard, the system is committed to ongoing improvements in accuracy, ensuring that the extracted data remain robust and responsive to emerging trends (Nagy & Reis, 2023). These quality-focused advances are especially relevant given the substantial changes observed in labour markets between 2017 and 2021, largely influenced by the rapid digital transformation accelerated by the COVID-19 pandemic (see also Cedefop, 2021, 2022).

The landscaping exercise found that the expansion in the OJA market has also triggered notable shifts in how PESs operate. Positioned as key portals for job postings, PESs are now more than job aggregators; they also facilitate AI-driven job-matching services, helping employers and jobseekers connect efficiently. This role is crucial for supporting disadvantaged jobseekers and ensuring greater inclusivity in the job market. PES portals provide free matching tools that are often chargeable on private job portals, reinforcing PESs' pivotal role in the evolving digital recruitment ecosystem. As the number of OJA operators grows, with the increasing presence of multinational platforms and a trend towards market consolidation, PESs' centralised role highlights the value of accessible, reliable recruitment resources.

The study also explores the application of skills intelligence from OJA data that may inform the ESCO framework, particularly within digital or green skills. Using LLMs, the study was able to sort and categorise terms into relevant fields, such as specific devices, programming languages and tools, ultimately proposing

an updated set of digital skills terms for the ESCO framework. This showcases how AI-driven classification can streamline and improve skills ontologies, ensuring they keep up with technological advances. However, challenges persist concerning green skills in OJAs, where 'green by definition' occupations like environmental engineering often omit specific green skills in job requirements. To address this, an innovative text-filtering technique was introduced to distinguish skill-related content from company mission statements, enhancing the relevance of the green skills extracted.

An ongoing challenge is the need for cross-country comparability in skills intelligence, particularly in the multilingual context of European labour markets. Differences in language and terminology and cultural nuances make creating universal, cross-national skills categorisations complex. A hybrid approach, blending a computational approach with human oversight and validation, is essential to ensure that ontologies are current and adaptable to linguistic diversity. Additionally, establishing a close partnership with the ESCO team would help synchronise updates with real-world labour trends and multilingual requirements, making the ESCO framework more attuned to evolving demands for skills.

As demonstrated throughout this study, the DPS framework's analysis of OJAs provides foundational insights for labour market analysis. Firstly, the detailed skills data extracted from OJAs allows a more granular understanding of job market demand, including precise information about the certificates and qualifications required. Secondly, the DPS facilitates an in-depth view of labour market dynamics, shedding light on regional disparities in skills demand. Thirdly, stakeholders can pinpoint skills gaps by integrating DPS data with workforce supply information, enabling targeted skills development and thus supporting programmes and training initiatives. (e.g. Solas, 2023) In summary, this data-driven approach to skills intelligence holds transformative potential for shaping labour market strategies, empowering stakeholders – including employers, educators and policymakers – to respond effectively to the demands of an increasingly digital and environmentally focused economy.

In moving forward, further refinement of the data-driven approach is recommended. Creating a training dataset annotated by experts could improve the accuracy of skills extraction, particularly in fields of study and cross-national contexts where linguistic and cultural factors introduce variability. By enriching the dataset with expert-verified annotations, the DPS could enhance its machine learning capabilities, ultimately providing a more robust and adaptable foundation for ongoing analysis of green and digital skills in the OJA landscape. Overall, the study underscores the importance of continued innovation and collaboration in

developing a resilient skills intelligence system that can adapt swiftly to the rapid shifts in the global labour market and support a future-ready workforce.

Abbreviations

AHP	analytical hierarchy process
AI	artificial intelligence
API	application programming interface
DPS	data production system
EFTA	European Free Trade Association
ESCO	European Skills, Competences, Qualifications and Occupations
ESS	European Statistical System
ESSnet	Network of European Statistical System
EU-27	27 Member States of the EU
ICE	international country expert
ICT	information and communications technology
ISCED	International Standard Classification of Education
ISCED-F	International Standard Classification of Education – Fields of education and training
ISCO/ ISCO-08	International Standard Classification of Occupations 2008
IT	information technology
LLM	large language model
NSI	national statistical institute
OJA	online job advertisement
PES	public employment service
UNESCO	United Nations Educational, Scientific and Cultural Organisation
WIH	Web Intelligence Hub

References

[All URLs accessed 11 February 2025]

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics, 40*(S1), S293-S340. https://doi.org/10.1086/718327

Auktor, G. (2021). *Green industrial skills for a sustainable future*. United Nations Industrial Development Organization. www.unido.org/sites/default/files/files/2021-02/LKDForum-2020_Green-Skills-for-a-Sustainable-Future.pdf

Ascheri, A., Kiss Nagy, A., Marconi, G., et al. (2022). *Competition in urban hiring markets: Evidence from online job advertisements – 2021 edition.* Luxembourg: Publications Office of the European Union. https://data.europa.eu/doi/10.2785/667004

Babina, T., Fedyk, A., He, A. X., & Hodson, J. (2023). Firm investments in artificial intelligence technologies and changes in workforce. SSRN. http://dx.doi.org/10.2139/ssrn.4060233

Bennett, F., Escudero, V., Liepmann, H., & Podjanin, H. (2022). Using online vacancy and job applicants' data to study skills dynamics. *IZA discussion paper No. 15506*. https://docs.iza.org/dp15506.pdf

Bowen, A., Kuralbayeva, K., & Tipoe, E. L. (2018). Characterising green employment: The impacts of 'greening' on workforce composition. *Energy Economics, 72*, 263–275. https://doi.org/10.1016/j.eneco.2018.03.015

Cedefop. (2008). *Terminology of European education and training policy*. www.cedefop.europa.eu/en/tools/vet-glossary/glossary

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

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

Cedefop. (2021). *Trends, transitions, and transformation*. Luxembourg: Publications Office of the European Union. Cedefop briefing note, April 2021. http://data.europa.eu/doi/10.2801/33991

Cedefop. (2022). *Employment trends during the Covid-19 pandemic: Skills intelligence data insight*. www.cedefop.europa.eu/en/data-insights/employment-trends-during-covid-19-pandemic

Cedefop. (2023a). *Skills in transition: The way to 2035*. Luxembourg: Publications Office of the European Union. http://data.europa.eu/doi/10.2801/438491

Cedefop. (2023b). *Going digital means skilling for digital: Using big data to track emerging digital skill needs*. Luxembourg: Publications Office of the European Union. http://data.europa.eu/doi/10.2801/772175

Cedefop. (2024). *Tracking the green transition in labour markets: Using big data to identify the skills that make jobs greener*. Publications Office of the European Union. Cedefop policy brief. www.cedefop.europa.eu/en/publications/9197

Chaise. (2021). *Study on blockchain labour market characteristics: A blueprint for sectoral cooperation on blockchain skill development. Deliverable D2.2.1*. https://chaise-blockchainskills.eu/wp-content/uploads/2021/05/D2.2.1-Study-on-Blockchain-labour-market-characteristics.pdf

Chen, CW, Li, LY. (2023). Is hiring fast a good sign? Is hiring fast a good sign? The informativeness of job vacancy duration for future frm proftability, *Review of Accounting Studies,* 28, 1316-1353. https://link.springer.com/article/10.1007/s11142-023-09797-2

Consoli, D., Marin, G., Marzucchi, A., & Vona, F. (2016). Do green jobs differ from non-green jobs in terms of skills and human capital? *Research Policy, 45*(5), 1046–1060. https://doi.org/10.1016/j.respol.2016.02.007

Dawson, N., Molitorisz, S., Rizoiu, M.-A., & Fray, P. (2021). Layoffs, inequity, and COVID-19: A longitudinal study of the journalism jobs crisis in Australia from 2012 to 2020. *Journalism, 24*(3). https://doi.org/10.1177/1464884921996286

Descy, P., Kvetan, V., Wirthmann, A., & Reis, F. (2019). Towards a shared infrastructure for online job advertisement data. *Statistical Journal of the IAOS, 1*, 669-675. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190547

Dierdorff, E. C., Norton, J. J., Drewes, D. W., Rivkin, D., & Lewis, P. (2009). *Greening of the world of work: Implications for O*NET®-SOC and new and emerging occupations*. National Center for O*NET Development. www.onetcenter.org/dl_files/Green.pdf

Fabo, B., Beblavý, M., & Lenaerts, K. (2017). The importance of foreign language skills in the labour markets of Central and Eastern Europe: Assessment based on data from online job portals. *Empirica, 44*(3), 487-508. https://doi.org/10.1007/s10663-017-9374-6

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2022). Embeddings Evaluation Using a Novel Measure of Semantic Similarity. *Cognitive Computation*, *14*(2), 749-763.

Grüger, J., & Schneider, G. (2019). Automated analysis of job requirements for computer scientists in online job advertisements. *Proceedings of the 15th International Conference on Web Information Systems and Technologies*.

Leigh, N. G., Lee, H., & Kraft, B. (2020). Robots, skill demand, and manufacturing in US regional labour markets. *Cambridge Journal of Regions, Economy and Society, 13*(1), 77-97. https://doi.org/10.1093/cjres/rsz019

Macedo, M. M., Clarke, W., Lucherini, E., Baldwin, T. Queiroz Neto, D., et al. (2022). Practical skills demand forecasting via representation learning of temporal dynamics. https://arxiv.org/pdf/2205.09508v1

Marrero-Rodríguez, R., Morini-Marrero, S., & Ramos-Henriquez, J. M. (2020). Tourism jobs in demand: Where the best contracts and high salaries go at online offers. *Tourism Management Perspectives, 35*, 100721. https://doi.org/10.1016/j.tmp.2020.100721

Muennighoff, Tazi, N., Magni, L. & Reimers, N. (2023). MTEB: Massive Text Embedding Benchmark. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2014–2037). https://doi.org/10.18653/v1/2023.eacl-main.148

Nagy, A.-M., & Reis, F. (2023). Development of an OJA gold standard to classify occupation. *Conference on New Techniques and Technologies for Statistics (NTTS) – Brussels*.

Nagy, A.-M., Gotuzzo, E., & Reis, F. (2024). Innovative approaches to enhance data quality in official statistics: A case study on online job advertisement data. *11th European Conference on Quality in Official Statistics (Q2024)*.

Napierala, J., Kvetan, V., & Branka, J. (2022). *Assessing the representativeness of online job advertisements*. Luxembourg: Publications Office of the European Union. Cedefop working paper No. 17. http://data.europa.eu/doi/10.2801/807500

Napierala, J. (2024). Enhancing taxonomy-based extraction: Leveraging information from online community platforms for digital skills demand identification in job ads. *Statistical Journal of the IAOS, 1*, 1-12. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji230110

Nasir, S. A. M., Wan Yaacob, W. F., & Wan Aziz, W. A. H. (2020). Analysing online vacancy and skills demand using text mining. *Journal of Physics: Conference Series* (1496). doi:https://doi.org/10.1088/1742-6596/1496/1/012011

NCVER. (2005). *Qualifications use for recruitment in the Australian labour market*. www.ncver.edu.au/__data/assets/file/0018/4536/nr1021.pdf

OECD (2023), *Big Data Intelligence on Skills Demand and Training in Umbria*, OECD Publishing, Paris, www.oecd.org/en/publications/big-data-intelligence-on-skills-demand-and-training-in-umbria_4bbbbfd6-en.html

Pater, R., Szkola, J., & Kozak, M. (2019). A method for measuring detailed demand for workers' competences. *Economics, 13*(1). https://doi.org/10.5018/economics-ejournal.ja.2019-27

Prüfer, J., & Prüfer, P. (2019). Data science for entrepreneurship research: Studying demand dynamics for entrepreneurial skills in the Netherlands. *Small Business Economics, 55*(3), 651-672. https://doi.org/10.1007/s11187-019-00208-y

Saussay, A., Sato, M., Vona, F., & O'Kane, L. (2022). *Who's fit for the low-carbon transition? Emerging skills and wage gaps in job ad data.* Centre for Climate Change Economics and Policy working paper No. 406. www.lse.ac.uk/granthaminstitute/wp-content/uploads/2022/10/working-paper-381-Saussay-et-al.pdf

Solas. (2023). *Winter skills bulletin 2023: Transversal skills in Ireland's labour market (Q4 2022-Q3 2023)*. www.solas.ie/f/70398/x/9ef876b3f5/solas-winter-skills-bulletin.pdf

Ternikov, A., & Aleksandrova, E. (2020). Demand for skills on the labor market in the IT sector. *Business Informatics, 14*(2), 64-83. https://doi.org/10.17323/2587-814x.2020.2.64.83

Vona, F., Marin, G., Consoli, D., & Popp, D. (2018). *Journal of the Association of Environmental and Resource Economists, 5*(4), 713-753. www.journals.uchicago.edu/doi/abs/10.1086/698859

Vona, F., Marin, G., & Consoli, D. (2019). Measures, drivers, and effects of green employment: Evidence from US local labour markets, 2006-2014. *Journal of Economic Geography, 19*(5), 1021-1048. https://doi.org/10.1093/jeg/lby038

Vona, F. (2021). *Labour markets and the green transition: A practitioner's guide to the task-based approach*. F. Biagi & A. Bitat (Eds.). Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/65924

Watts, R. D., Bowles, D. C., Fisher, C., & Li, I. W. (2019). Public health job advertisements in Australia and New Zealand: A changing landscape. *Australian and New Zealand Journal of Public Health*, 43(6), 522-528. https://doi.org/10.1111/1753-6405.12931

Annexes

Annex 1.
Infrastructure of the online job advertisement data production system

Hardware components

(a) EC2 instances. The DPS relies on Elastic Compute Cloud (EC2) instances to host various components such as Xdiana, MongoDB and AWS EMR clusters. EC2 instances provide scalable computing capacity to accommodate fluctuating workloads.

(b) r5.12xlarge instances. Specifically, AWS EMR clusters employ instances like r5.12xlarge for processing large volumes of data. These instances offer high-performance computing capabilities essential for data analysis tasks.

Software components

Xdiana. The data ingestion module is powered by Xdiana, which is responsible for collecting, crawling and scraping data from online sources. Xdiana is hosted on an EC2 instance within the DPS infrastructure.

MongoDB. As an NoSQL database, MongoDB serves as the storage backend for the DPS, housing extracted data, configuration information, web pages and job postings. MongoDB's flexibility and scalability support the diverse data storage needs of the system.

Apache Spark. For data processing tasks, the DPS utilises Apache Spark, an open-source distributed computing framework. Spark enables parallel processing of large datasets across an AWS EMR cluster, ensuring efficient analysis and machine learning application.

AWS services. The DPS leverages various AWS services for orchestration, storage and backup, including AWS S3 for object storage, AWS Step Functions for workflow orchestration and AWS Glacier for backup and archival purposes.

AWS Athena. The analytical layer of the DPS interfaces with AWS Athena, providing users with SQL-like access to data and supporting external connections for analysis and reporting.

Network components

AWS infrastructure. The entire DPS infrastructure operates within the AWS cloud environment, leveraging AWS's global network infrastructure for reliable and scalable network connectivity.

Organisational structure

Orchestration. Workflow orchestration within the DPS is managed by AWS Step Functions, which automates the provisioning and management of resources, including EC2 instances and EMR clusters.

Data management. Data management tasks, including ingestion, processing, storage and backup, are coordinated within a structured workflow to ensure data integrity, reliability and accessibility.

Annex 2.
Detailed tables for chapter 5

Table 6. **List of selected digital occupations**

ISCO code	Occupation description
1330	ICT service managers
2511	Systems analysts
2512	Software developers
2513	Web and multimedia developers
2514	Applications programmers
2519	Software and applications developers and analysts
2521	Database designers and administrators
2522	Systems administrators
2523	Computer network professionals
2529	Database and network professionals
3511	ICT operations technicians
3512	ICT user support technicians
3513	Computer network and systems technicians
3514	Web technicians
3521	Broadcasting and audiovisual technicians

Table 7. **Examples of terms with probability scores across category groups**

term	Programming language	Operating system	Machine learning algorithm	Data base	Software framework	Digital device	Computer network	Cloud computing	Computer programming keyword	IoT	Software tool	Computer hardware	File format	Best_ labels	N_top_ labels
cxx	1	0,8	0,62	0,56	0,56	0,55	0,53	0,39	0,33	0,33	0,32	0,01	0	['programming language']	1
file	0	0,17	0,2	0,25	0,24	0,9	0,58	0,73	0,06	0,31	0,29	0,04	0,78	['digital device']	1
windows - phone - 8.1	0,01	1	0,24	0,02	0,94	0,99	0,02	0,04	0,88	0	0,97	0	0,09	['operating system', 'software framework', 'digital device', 'software tool']	4
star - schema	0,01	0,04	0,06	0,98	0,27	0,04	0,26	0,07	0,09	0,16	0,64	0,01	0,12	['database']	1
compass - geolocation	0	0,01	0,02	0	0,09	0,97	0,18	0,01	0,11	0,4	0,51	0,01	0,02	[digital device']	1
tensorflow	0,97	0	0,99	0,23	0,99	0,49	0,81	0,58	0,77	0,47	0,99	0	0,85	['programming language', 'machine learning algorithm', 'software framework', 'software tool']	4
redshift	0,1	0,21	0,04	0,99	0,37	0,07	0,94	0,52	0,59	0,26	0,99	0,45	0,19	['database', 'computer network', 'software tool']	3
websecurity	0,24	0,11	0,1	0,34	0,3	0,04	0,95	0,22	0,3	0,23	0,93	0	0	['computer network', 'software tool']	2
artificial - intelligence	0,07	0,11	0,93	0,16	0,42	0,65	0,36	0,07	0,09	0,25	0,78	0,01	0,28	['machine learning algorithm']	1

IoT = internet of things.
Source: Authors.

Annex 3.
Step-by-step description of the bottom-up data driven to green skills and the list of green skills terms

The approach was developed in two phases. The first phase consisted of the following steps.

(a) Enhancement of green terms using OJAs. Using an exact match, filtered OJAs with 141 green terms focusing on green-related jobs were used to refine broad terms (e.g. ‘eco’ to ‘eco-friendly’) and expand narrower ones (e.g. ‘air quality’ to ‘air pollution’).

(b) Iterative term enrichment. Using the corpus of filtered OJAs and a fast text model (obtained from Giabelli et al., 2022), we extracted the most similar words (the measure used was cosine similarity). Afterwards, we selected the 10 words most similar (note that we did not choose any specific threshold here) and then cleaned the results) to the initial 141 green terms to expand the green terms from 141 to 227 by adding lexicon variations (e.g. ‘sustainability’ to ‘sustainability agenda’ or ‘low carbon’ to ‘low carbon economy’), reducing noise and adding relevant vacancies.

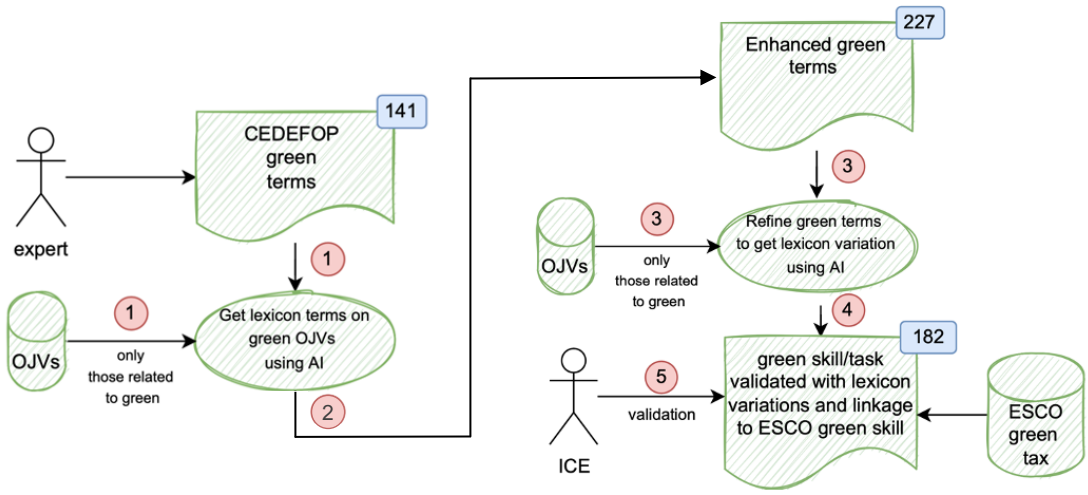
(c) Linguistic variation identification. The model was further enhanced by identifying linguistic variations used by employers, improving its ability to classify green-related jobs.

(d) Top mentions identification. We refined the model to identify the top four linguistic variations (synonyms, specifications or generalisations) for each green term as they emerged in OJAs. We extracted portions of OJA text that contained the initial Cedefop words to find a more precise or complete mention; for example, when looking for ‘air hygiene’, we extracted ‘air hygiene management’.

(e) Human validation and translation. We engaged human experts (ICEs) to validate the skill taxonomy by distinguishing tasks from skills and ensuring linguistic coherence. We then translated mentions into four languages (Dutch, French, German and Italian), refining translations for clarity and consistency.

Finally, the linkage to the ESCO classification was created by looking again for the most similar terms (top five in terms of cosine similarity). Figure 34 illustrates the process with the corresponding number of skills obtained from each step.

Figure 34. **Process for obtaining green skills in the first phase**



tax = taxonomy.

Source: Authors.

The second phase consisted of the following steps:

(a) Validation. The ICEs validated the green terms and skills obtained from the first project in all official EU languages.

(b) Filtering and matching. OJAs were filtered for green terms/skills using a word embedding pipeline, with a threshold of cosine similarity of 1 to ensure precision and reduce false positives.

(c) Indicators. Two indicators were created:

  (i) greenness: ratio of green skills to generic skills in OJAs;

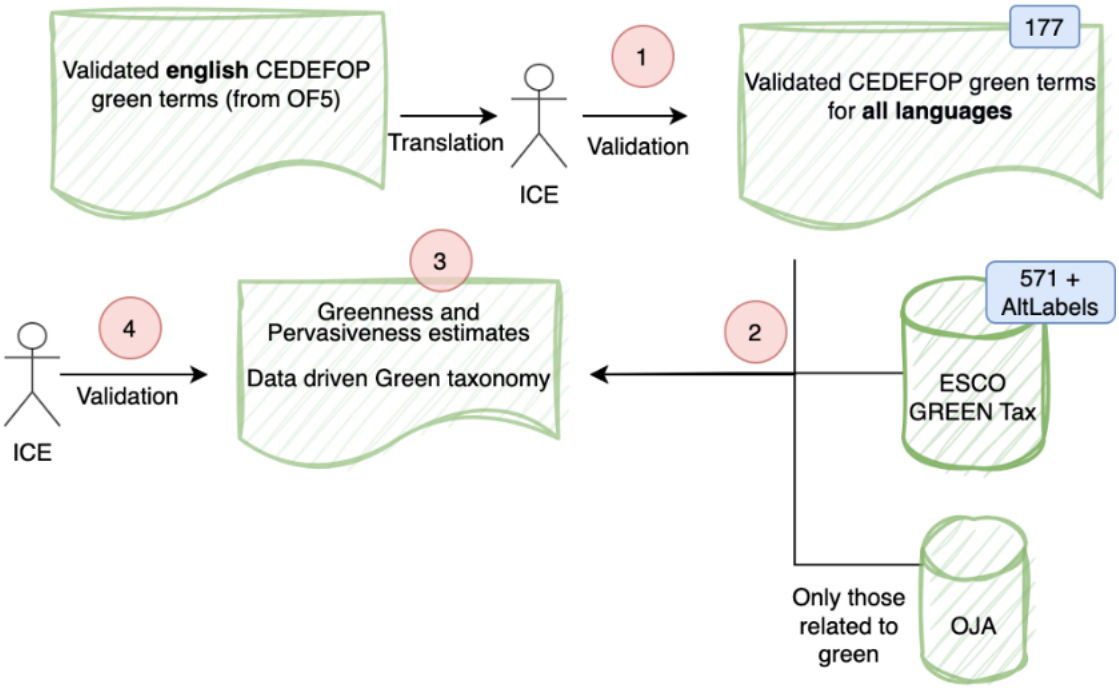
  (ii) green pervasiveness: ratio of green OJAs to total OJAs for each occupation; occupations were selected if they had 1% or more pervasiveness and two or more green skills.

(d) Human validation. ICEs verified the coherence of green skills within their associated occupations.

Figure 35 shows the procedure implemented in the second phase. Note that the number of validated terms is now 177 and not 182, as some were excluded due to their scarce relevance or excessive generality (e.g. sustainable).

Annex 3.
Step-by-step description of the bottom-up data driven to green skills and the list of green skills terms

Figure 35. **Workflow process in the second phase**



NB: AltLabels = alternative labels; OF5 = step 5 in Figure 34; tax = taxonomy.

Source: Authors.

Table 8. **Bag of green terms together with enhanced mentions observed in OJAs and associated green ESCO terms**

Green term	Green mention	Top 3 associated ESCO green terms
air hygiene	air hygiene management	['(perform water treatment – 52.4)', '(carpooling service – 42.3)']
air hygiene	provide air hygiene services	['(perform water treatment – 52.4)', '(carpooling service – 42.3)']
air pollution	measuring of air pollution	['(measure pollution – 72.5)', '(prevent marine pollution – 65.8)', '(prevent sea pollution – 65.6)']
air pollution	air pollution monitoring	['(measure pollution – 72.5)', '(prevent marine pollution – 65.8)', '(prevent sea pollution – 65.6)']
air quality	air quality monitoring	['(environmental indoor quality – 74.0)', '(monitor water quality – 53.9)', '(plant contribution indoor climate health condition – 50.0)']
air quality	carrying out an air quality survey	['(environmental indoor quality – 74.0)', '(monitor water quality – 53.9)', '(plant contribution indoor climate health condition – 50.0)']
assessment drainage	environmental impact assessment and drainage strategy	['(design drainage well system – 69.4)']
assessment drainage	delivering flood risk assessments and drainage strategies	['(design drainage well system – 69.4)']
carbon emission	reduce carbon emission	['(reduce tanning emission – 51.5)', '(energy efficiency – 51.0)']
carbon footprint	reducing carbon footprint	['(analyse energy consumption – 51.9)', '(gas consumption – 45.5)']
chemistry environmental	chemistry and environmental engineering	['(soil science – 65.7)', '(geology – 63.7)', '(pest biology – 58.2)']
chemistry environmental	chemistry and environmental science	['(soil science – 65.7)', '(geology – 63.7)', '(pest biology – 58.2)']

Green term	Green mention	Top 3 associated ESCO green terms
contamination assessment	contamination assessment and determining the remedial strategies	['(prepared animal feed contamination hazard – 69.4)', '(ass contamination – 68.2)', '(avoid contamination – 60.8)']
contamination assessment	hydrogeologic contamination assessment and determining the remedial strategies	['(prepared animal feed contamination hazard – 69.4)', '(ass contamination – 68.2)', '(avoid contamination – 60.8)']
contamination assessment	land contamination assessment	['(prepared animal feed contamination hazard – 69.4)', '(ass contamination – 68.2)', '(avoid contamination – 60.8)']
contamination assessment	reporting contamination assessment	['(prepared animal feed contamination hazard – 69.4)', '(ass contamination – 68.2)', '(avoid contamination – 60.8)']
corporate sustainability	measuring of corporate sustainability	['(promote sustainability – 81.1)', '(measure sustainability tourism activity – 72.7)', '(advise sustainability solution – 68.7)']
disposal waste	disposal waste collection	['(handle waste – 60.7)', '(hazardous waste storage – 59.5)', '(educate hazardous waste – 56.8)']
disposal waste	disposal waste removal	['(handle waste – 60.7)', '(hazardous waste storage – 59.5)', '(educate hazardous waste – 56.8)']
district energy	skills in district energy engineering	['(integrate biogas energy building – 59.6)', '(district heating cooling – 58.1)', '(energy performance building – 53.8)']
district heating	district heating maintenance	['(solar thermal energy system hot water heating – 53.5)', '(perform feasibility study solar heating – 52.4)', '(perform feasibility study electric heating – 52.0)']
drainage area	completing drainage area	['(design drainage well system – 65.6)']
drainage area	drainage area planning	['(design drainage well system – 65.6)']
drainage flood	drainage flood	['(design drainage well system – 70.4)']
drainage infrastructure	designing of drainage infrastructure	['(design drainage well system – 73.9)', '(promote innovative infrastructure design – 60.7)', '(watershed development – 52.9)']
drainage strategy	detailed drainage strategy	['(design drainage well system – 75.8)', '(green space strategy – 48.1)']

Green term	Green mention	Top 3 associated ESCO green terms
drainage strategy	foul drainage strategy	['(design drainage well system – 75.8)', '(green space strategy – 48.1)']
drainage strategy	strategies of surface water drainage	['(design drainage well system – 75.8)', '(green space strategy – 48.1)']
earthwork remediation	earthwork remediation	['(develop bioremediation technique – 60.8)', '(develop flood remediation strategy – 54.2)', '(develop site remediation strategy – 50.1)']
ecological assessment	ecological assessment for landscape planning	['(ecological principle – 77.2)', '(conduct ecological research – 71.8)', '(conduct ecological survey – 68.6)']
ecological assessment	undertaking ecological assessment	['(ecological principle – 77.2)', '(conduct ecological research – 71.8)', '(conduct ecological survey – 68.6)']
ecological assessment	preliminary ecological assessment	['(ecological principle – 77.2)', '(conduct ecological research – 71.8)', '(conduct ecological survey – 68.6)']
ecological assessment	specialised ecological assessment	['(ecological principle – 77.2)', '(conduct ecological research – 71.8)', '(conduct ecological survey – 68.6)']
ecological constraint	ecological constraint project	['(conduct ecological survey – 59.3)', '(plant specie – 58.9)', '(ensure safety endangered specie protected area – 51.8)']
ecological constraint	identifying ecological constraint	['(conduct ecological survey – 59.3)', '(plant specie – 58.9)', '(ensure safety endangered specie protected area – 51.8)']
ecological consultancy	deliver ecological consultancy	['(ecology – 74.5)', '(ecological principle – 66.4)', '(agroecology – 64.9)']
ecological consultancy	deliver ecological consultancy and mitigation works	['(ecology – 74.5)', '(ecological principle – 66.4)', '(agroecology – 64.9)']
ecological consultancy	experience within ecological consultancy	['(ecology – 74.5)', '(ecological principle – 66.4)', '(agroecology – 64.9)']
ecological consultant	consultancy of ecological and environment issues	['(ecology – 72.0)', '(agroecology – 63.5)', '(ecological principle – 62.4)']
ecological consultant	ecological consultancy	['(ecology – 72.0)', '(agroecology – 63.5)', '(ecological principle – 62.4)']
ecological contracting	creating ecological contracting	['(manage aquatic habitat – 54.4)', '(manage habitat – 50.4)', '(evaluate vehicle ecological footprint – 47.6)']

Green term	Green mention	Top 3 associated ESCO green terms
ecological impact	ecological impact appraisal	['(conduct ecological research – 72.7)', '(conduct ecological survey – 71.6)', '(ass environmental impact – 68.7)']
ecological impact	ecological impact assessment	['(conduct ecological research – 72.7)', '(conduct ecological survey – 71.6)', '(ass environmental impact – 68.7)']
ecological impact	developing solutions for ecological impact	['(conduct ecological research – 72.7)', '(conduct ecological survey – 71.6)', '(ass environmental impact – 68.7)']
ecological impact	undertaking ecological impact assessment	['(conduct ecological research – 72.7)', '(conduct ecological survey – 71.6)', '(ass environmental impact – 68.7)']
ecological issue	support ecological issue	['(ecological principle – 62.5)', '(conduct ecological research – 59.6)', '(conduct ecological survey – 54.7)']
ecological mitigation	ecological mitigation completion	['(conduct ecological survey – 62.3)', '(conduct ecological research – 62.1)', '(analyse ecological data – 60.6)']
ecological mitigation	ecological mitigation work	['(conduct ecological survey – 62.3)', '(conduct ecological research – 62.1)', '(analyse ecological data – 60.6)']
ecological mitigation	implementing ecological mitigation	['(conduct ecological survey – 62.3)', '(conduct ecological research – 62.1)', '(analyse ecological data – 60.6)']
ecological project	development of ecological project	['(ecological principle – 79.1)', '(conduct ecological survey – 63.2)', '(conduct ecological research – 62.9)']
ecological project	ecological project management	['(ecological principle – 79.1)', '(conduct ecological survey – 63.2)', '(conduct ecological research – 62.9)']
ecological project	managing ecological project	['(ecological principle – 79.1)', '(conduct ecological survey – 63.2)', '(conduct ecological research – 62.9)']
ecological specialism	licence for ecological specialisation	['(plant specie – 53.9)', '(botany – 41.5)']
ecological survey	assessment of ecological surveys	['(ecological principle – 74.2)', '(conduct ecological research – 71.0)', '(conduct environmental survey – 63.5)']
ecological survey	carry out ecological surveys	['(ecological principle – 74.2)', '(conduct ecological research – 71.0)', '(conduct environmental survey – 63.5)']

Green term	Green mention	Top 3 associated ESCO green terms
ecological survey	wide range of ecological surveys	['(ecological principle – 74.2)', '(conduct ecological research – 71.0)', '(conduct environmental survey – 63.5)']
ecological survey	undertaking ecological survey	['(ecological principle – 74.2)', '(conduct ecological research – 71.0)', '(conduct environmental survey – 63.5)']
ecology consultancy	ecology consultancy	['(ecology – 68.6)', '(wildlife project – 51.0)', '(urban planning law – 46.8)']
ecology consultancy	tracking record in ecology consultancy	['(ecology – 68.6)', '(wildlife project – 51.0)', '(urban planning law – 46.8)']
electric car	diagnostics of electric car	['(electric motor – 70.4)', '(electric heating system – 63.4)', '(describe electric drive system – 47.8)']
electric car	electric car charging	['(electric motor – 70.4)', '(electric heating system – 63.4)', '(describe electric drive system – 47.8)']
energy conservation	programming energy conservation	['(energy efficiency – 62.4)', '(energy performance building – 58.2)', '(analyse energy consumption – 54.1)']
energy conservation	implementation of strategies for energy conservation	['(energy efficiency – 62.4)', '(energy performance building – 58.2)', '(analyse energy consumption – 54.1)']
energy consumption	energy consumption designing	['(advise utility consumption – 77.1)', '(gas consumption – 76.6)', '(water consumption – 69.9)']
energy consumption	energy consumption maintenance	['(advise utility consumption – 77.1)', '(gas consumption – 76.6)', '(water consumption – 69.9)']
energy reduction	delivering energy reduction projects	['(energy efficiency – 74.0)', '(advise heating system energy efficiency – 64.3)', '(design indicator food waste reduction – 63.9)']
energy sustainability	energy sustainability management	['(advise sustainability solution – 69.6)', '(energy performance building – 65.3)', '(promote sustainability – 61.0)']
energy sustainable development	delivering energy sustainable development	['(sustainable development goal – 78.4)', '(provide training sustainable tourism development management – 64.6)', '(promote use sustainable transport – 60.6)']
energy transition	energy transition service	['(climate change impact – 52.1)', '(green bond – 41.1)']

Green term	Green mention	Top 3 associated ESCO green terms
environmental appraisal	environmental appraisal	['(conduct environmental site assessment – 66.4)', '(environmental impact tourism – 65.7)', '(communicate environmental impact mining – 65.6)']
environmental appraisal	planning of environmental appraisal	['(conduct environmental site assessment – 66.4)', '(environmental impact tourism – 65.7)', '(communicate environmental impact mining – 65.6)']
environmental appraisal	developing environmental statement and environmental appraisal	['(conduct environmental site assessment – 66.4)', '(environmental impact tourism – 65.7)', '(communicate environmental impact mining – 65.6)']
environmental aspect	identifying environmental aspects of project's activities	['(promote environmental awareness – 71.5)', '(implement environmental action plan – 71.3)', '(airport environmental regulation – 69.5)']
environmental aspect	environmental aspect management	['(promote environmental awareness – 71.5)', '(implement environmental action plan – 71.3)', '(airport environmental regulation – 69.5)']
environmental assessment	management of environmental assessments	['(ass environmental impact – 71.3)', '(manage environmental impact – 68.8)', '(environmental impact tourism – 68.2)']
environmental assessment	environmental assessment method	['(ass environmental impact – 71.3)', '(manage environmental impact – 68.8)', '(environmental impact tourism – 68.2)']
environmental assessment	strategic environmental assessment	['(ass environmental impact – 71.3)', '(manage environmental impact – 68.8)', '(environmental impact tourism – 68.2)']
environmental audit	carry out environmental audit	['(environmental management monitor – 77.1)', '(implement environmental action plan – 72.5)', '(monitor farm environmental management plan – 72.1)']
environmental audit	environmental audit inspection	['(environmental management monitor – 77.1)', '(implement environmental action plan – 72.5)', '(monitor farm environmental management plan – 72.1)']
environmental data	review of environmental data	['(environmental indoor quality – 70.6)', '(analyse ecological data – 69.4)', '(collect biological data – 69.0)']
environmental hazard	ensure environmental hazard reporting	['(environmental threat – 69.4)', '(environmental impact tourism – 62.0)', '(environmental aspect inland waterway transportation – 61.0)']

Green term	Green mention	Top 3 associated ESCO green terms
environmental hazard	prepare reports concerning environmental hazard	['(environmental threat – 69.4)', '(environmental impact tourism – 62.0)', '(environmental aspect inland waterway transportation – 61.0)']
environmental hazard	safety and environmental hazard identification	['(environmental threat – 69.4)', '(environmental impact tourism – 62.0)', '(environmental aspect inland waterway transportation – 61.0)']
environmental issue	managing safety environmental issues	['(environmental threat – 70.8)', '(advise environmental risk management system – 68.0)', '(carry training environmental matter – 67.2)']
environmental management	management of environmental safety	['(implement environmental action plan – 80.4)', '(environmental indoor quality – 77.6)', '(environmental legislation – 74.3)']
environmental monitoring	undertaking environmental monitoring and sampling activities	['(food waste monitoring system – 67.9)', '(perform environmental investigation – 65.8)', '(perform environmental remediation – 64.8)']
environmental monitoring	undertaking environmental monitoring	['(food waste monitoring system – 67.9)', '(perform environmental investigation – 65.8)', '(perform environmental remediation – 64.8)']
environmental performance	environmental performance goal reaching	['(measure company's sustainability performance – 73.7)', '(environmental management monitor – 72.8)', '(environmental indoor quality – 71.0)']
environmental performance	ensuring safety and environmental performance	['(measure company's sustainability performance – 73.7)', '(environmental management monitor – 72.8)', '(environmental indoor quality – 71.0)']
environmental permit	managing environmental permits	['(environmental management monitor – 70.3)', '(monitor farm environmental management plan – 64.6)', '(pollution legislation – 64.2)']
environmental protection	ensuring safety and environmental protection	['(environmental policy – 67.1)', '(advise soil water protection – 63.3)', '(develop environmental policy – 62.6)']
environmental quality	ensuring safety and environmental quality	['(environmental policy – 70.8)', '(ICT environmental policy – 62.3)', '(develop environmental policy – 61.2)']
environmental regulation	ensuring environmental regulation standards	['(health safety regulation – 82.5)', '(environmental legislation – 76.6)', '(asbestos removal regulation – 75.2)']
environmental science	environmental science and chemistry	['(soil science – 79.1)', '(Earth science – 77.3)', '(environmental engineering – 70.7)']

Green term	Green mention	Top 3 associated ESCO green terms
environmental science	environmental science engineering	['(soil science – 79.1)', '(Earth science – 77.3)', '(environmental engineering – 70.7)']
environmental science	qualification in environmental science	['(soil science – 79.1)', '(Earth science – 77.3)', '(environmental engineering – 70.7)']
environmental science	relevant environmental science experience	['(soil science – 79.1)', '(Earth science – 77.3)', '(environmental engineering – 70.7)']
environmental service	engineering environmental service	['(environmental aspect inland waterway transportation – 61.0)', '(manage air quality – 53.2)', '(train staff waste management – 51.3)']
environmental service	environmental service responsibility	['(environmental aspect inland waterway transportation – 61.0)', '(manage air quality – 53.2)', '(train staff waste management – 51.3)']
environmental service	delivering a range of environmental services	['(environmental aspect inland waterway transportation – 61.0)', '(manage air quality – 53.2)', '(train staff waste management – 51.3)']
environmental specialist	regeneration environmental specialisation	['(coordinate environmental effort – 63.6)', '(train staff waste management – 52.4)']
environmental standard	standard environmental skill	['(ensure compliance environmental legislation food production – 75.6)', '(ensure compliance environmental legislation – 74.7)', '(environmental policy – 72.1)']
environmental sustainability	developing MEP, environmental and sustainability services	['(advise sustainability solution – 76.5)', '(measure sustainability tourism activity – 75.9)', '(promote sustainability – 75.7)']
environmental waste	regulations relating to environmental and waste management	['(environmental threat – 71.1)', '(apply road transport environmental measure – 69.6)', '(manage environmental management system – 67.6)']
environmentally sustainable	designing of environmentally sustainable projects	['(apply sustainable tillage technique – 70.6)', '(promote use sustainable transport – 59.8)', '(promote sustainable energy – 57.2)']
flood risk	complete flood risk assessments	['(design drainage well system – 72.5)', '(conserve water resource – 55.5)', '(watershed development – 55.1)']
flood risk	flood risk assessment	['(design drainage well system – 72.5)', '(conserve water resource – 55.5)', '(watershed development – 55.1)']

Green term	Green mention	Top 3 associated ESCO green terms
flood risk	flood risk management	['(design drainage well system – 72.5)', '(conserve water resource – 55.5)', '(watershed development – 55.1)']
geo environmental	geo environmental engineering	['(perform environmental remediation – 68.5)', '(advise environmental remediation – 67.0)', '(conduct environmental survey – 61.6)']
geoenvironmental consultant	geoenvironmental consultancy BSc	['(treat contaminated water – 58.7)']
geoenvironmental consultant	geoenvironmental consultancy	['(treat contaminated water – 58.7)']
geoenvironmental consultant	licensed geoenvironmental consultancy	['(treat contaminated water – 58.7)']
geoenvironmental engineer	geoenvironmental engineering	['(environmental engineering – 69.7)', '(geology – 64.3)', '(treat contaminated water – 57.9)']
geoenvironmental engineer	geotechnical and geoenvironmental engineering	['(environmental engineering – 69.7)', '(geology – 64.3)', '(treat contaminated water – 57.9)']
geography environmental	geography and environmental science	['(environmental engineering – 69.5)', '(ass environmental impact – 57.6)', '(conduct airport environmental study – 57.6)']
geography environmental	geography and environmental studies	['(environmental engineering – 69.5)', '(ass environmental impact – 57.6)', '(conduct airport environmental study – 57.6)']
geology environmental	engineering and geology environmental sciences	['(geology – 81.0)', '(soil science – 67.0)', '(environmental engineering – 64.3)']
geology environmental	geology and environmental engineering	['(geology – 81.0)', '(soil science – 67.0)', '(environmental engineering – 64.3)']
geology environmental	geology and environmental science	['(geology – 81.0)', '(soil science – 67.0)', '(environmental engineering – 64.3)']
green energy	generation of green energy	['(renewable energy technology – 67.8)', '(solar energy – 53.5)', '(green bond – 53.4)']
heat pump	heat pump engineering	['(install solar water heater – 65.7)', '(solar thermal energy system hot water heating – 57.4)', '(domestic heating system – 52.1)']

Annex 3.
Step-by-step description of the bottom-up data driven to green skills and the list of green skills terms

Green term	Green mention	Top 3 associated ESCO green terms
innovative sustainable	designing innovative sustainable solutions	['(promote sustainable packaging – 67.6)', '(promote sustainable energy – 65.6)', '(promote sustainable interior design – 65.5)']
innovative sustainable	innovative and sustainable mechanical engineering	['(promote sustainable packaging – 67.6)', '(promote sustainable energy – 65.6)', '(promote sustainable interior design – 65.5)']
innovative sustainable	developing innovative sustainable solutions	['(promote sustainable packaging – 67.6)', '(promote sustainable energy – 65.6)', '(promote sustainable interior design – 65.5)']
land remediation	investigation on land remediation	['(perform environmental remediation – 67.0)', '(develop flood remediation strategy – 54.2)', '(develop bioremediation technique – 53.3)']
land remediation	land remediation consultancy	['(perform environmental remediation – 67.0)', '(develop flood remediation strategy – 54.2)', '(develop bioremediation technique – 53.3)']
landscape planning	landscape planning to reduce environmental impact	['(land use airport planning – 70.6)', '(promote innovative infrastructure design – 56.2)']
long sustainability	ensure long term sustainability	['(promote sustainability – 66.0)', '(provide training sustainable tourism development management – 62.5)', '(develop forestry strategy – 55.2)']
long sustainability	long term sustainability relationship	['(promote sustainability – 66.0)', '(provide training sustainable tourism development management – 62.5)', '(develop forestry strategy – 55.2)']
long sustainability	maintain oversight of the long term sustainability	['(promote sustainability – 66.0)', '(provide training sustainable tourism development management – 62.5)', '(develop forestry strategy – 55.2)']
low carbon economy	helping create a low carbon economy	['(ass life cycle resource – 44.1)']
low carbon technology	low carbon technology connection	['(ass hydrogen production technology – 57.4)', '(design smart grid – 47.0)', '(instruct energy saving technology – 41.7)']
low carbon technology	low carbon technology	['(ass hydrogen production technology – 57.4)', '(design smart grid – 47.0)', '(instruct energy saving technology – 41.7)']
marine environmental	marine environmental science	['(conduct environmental site assessment – 73.4)', '(ass environmental impact aquaculture operation – 71.7)', '(conduct environmental survey – 67.8)']

Green term	Green mention	Top 3 associated ESCO green terms
marine environmental	writing marine environmental statement	['(conduct environmental site assessment – 73.4)', '(ass environmental impact aquaculture operation – 71.7)', '(conduct environmental survey – 67.8)']
nature conservation	delivery of nature conservation	['(forest conservation – 76.1)', '(conservation agriculture – 66.2)', '(tree preservation conservation – 65.6)']
offshore wind	installation of offshore wind turbine	['(install onshore wind energy system – 63.8)', '(design offshore energy system – 60.9)', '(research location offshore farm – 59.9)']
onshore wind	engineering of onshore wind	['(research location offshore farm – 51.3)', '(type tidal stream generator – 48.7)', '(research location wind farm – 48.2)']
recycling facility	using recycling facility	['(monitor civic recycling site – 46.8)']
recycling plant	making recycling plant	['(handle mining plant waste – 64.0)', '(install recycling container – 61.4)', '(operate biogas plant – 60.3)']
reduce energy	proactively reduce energy	['(reduce tanning emission – 51.4)', '(monitor tree health – 41.8)', '(monitor forest health – 41.6)']
renewable energy	renewable energy project	['(solar energy – 65.6)', '(install onshore wind energy system – 61.7)', '(integrate biogas energy building – 61.1)']
sampling water	sampling water monitoring	['(manage water quality testing – 60.2)', '(monitor water quality – 57.9)', '(inspect water well – 54.0)']
sampling water	sampling water system	['(manage water quality testing – 60.2)', '(monitor water quality – 57.9)', '(inspect water well – 54.0)']
sampling water	experience of soil sampling and water monitoring	['(manage water quality testing – 60.2)', '(monitor water quality – 57.9)', '(inspect water well – 54.0)']
solar battery	designing solar battery project	['(solar energy – 58.5)', '(design solar energy system – 55.6)']
solar farm	solar farm portfolio	['(design wind farm collector system – 48.0)', '(solar panel mounting system – 41.6)']
solar panel	maintenance of solar panel	['(mount photovoltaic panel – 60.1)', '(type photovoltaic panel – 51.8)', '(design wind farm collector system – 42.4)']

Annex 3.
Step-by-step description of the bottom-up data driven to green skills and the list of green skills terms

Green term	Green mention	Top 3 associated ESCO green terms
solar panel	solar panel engineering	['(mount photovoltaic panel – 60.1)', '(type photovoltaic panel – 51.8)', '(design wind farm collector system – 42.4)']
surface water	assessment of surface water	['(model groundwater – 58.6)', '(study groundwater – 55.2)']
sustainability initiative	leading sustainability initiative	['(measure sustainability tourism activity – 73.5)', '(global standard sustainability reporting – 72.2)', '(measure company's sustainability performance – 65.9)']
sustainability issue	leadership of sustainability issues	['(advise sustainability solution – 71.2)', '(measure sustainability tourism activity – 60.2)', '(promote sustainability – 59.5)']
sustainability objective	achievement of sustainability objective	['(global standard sustainability reporting – 68.0)', '(implement sustainable procurement – 57.6)', '(develop food policy – 52.5)']
sustainability objective	requirement of sustainability objectives	['(global standard sustainability reporting – 68.0)', '(implement sustainable procurement – 57.6)', '(develop food policy – 52.5)']
sustainability objective	development of sustainability objective	['(global standard sustainability reporting – 68.0)', '(implement sustainable procurement – 57.6)', '(develop food policy – 52.5)']
sustainability objective	prepare a sustainability objective	['(global standard sustainability reporting – 68.0)', '(implement sustainable procurement – 57.6)', '(develop food policy – 52.5)']
sustainability policy	implementation of sustainability policies	['(develop food policy – 68.9)', '(develop energy policy – 68.7)', '(measure company's sustainability performance – 68.3)']
sustainability policy	understanding sustainability policies and processes	['(develop food policy – 68.9)', '(develop energy policy – 68.7)', '(measure company's sustainability performance – 68.3)']
sustainability principle	provide embed sustainability principles	['(crop production principle – 53.9)', '(fertilisation principle – 45.7)']
sustainable building	design of sustainable building	['(integrate biogas energy building – 67.8)', '(energy performance building – 64.8)', '(design domotic system building – 63.0)']
sustainable engineering	sustainable engineering	['(surface engineering – 64.1)', '(safety engineering – 63.9)', '(sustainable agricultural production principle – 53.9)']
sustainable procurement	maintain sustainable procurement strategies	['(sustainable forest management – 66.4)', '(promote sustainable packaging – 60.7)', '(develop wildlife program – 41.0)']

Green term	Green mention	Top 3 associated ESCO green terms
sustainable solution	develop sustainable solutions to problems	['(sustainable footwear material component – 60.1)', '(sustainable agricultural production principle – 55.8)', '(sustainable building material – 51.7)']
sustainable urban	designing of sustainable urban spaces	['(sustainable forest management – 63.9)', '(promote innovative infrastructure design – 54.5)', '(apply sustainable tillage technique – 46.4)']
treatment facility	design of treatment facilities	['(carry energy management facility – 69.2)']
waste collection	knowledge of waste collection	['(ass waste type – 57.0)', '(follow recycling collection schedule – 49.1)']
waste collection	packing waste collection	['(ass waste type – 57.0)', '(follow recycling collection schedule – 49.1)']
waste material	classification of waste material	['(dispose prepared animal feed waste – 65.2)', '(dispose food waste – 61.5)', '(dispose non-hazardous waste – 60.7)']
waste material	processing of waste material	['(dispose prepared animal feed waste – 65.2)', '(dispose food waste – 61.5)', '(dispose non-hazardous waste – 60.7)']
waste material	safety treatment of waste material	['(dispose prepared animal feed waste – 65.2)', '(dispose food waste – 61.5)', '(dispose non-hazardous waste – 60.7)']
waste stream	waste stream auditing	['(ass waste type – 70.8)', '(dispose waste – 67.5)', '(dispose food waste – 67.0)']
waste water treatment plant	maintenance of waste water treatment plant	['(carry waste water treatment – 91.1)', '(supervise waste water treatment – 86.5)', '(maintain water treatment equipment – 76.1)']
waste water treatment	wastewater treatment work	['(perform water treatment – 80.8)', '(maintain water treatment equipment – 70.2)', '(perform water treatment procedure – 67.7)']
water hygiene	water hygiene engineering	['(perform water treatment – 69.3)', '(carry waste water treatment – 59.5)', '(perform water treatment procedure – 58.4)']
water hygiene	technics of water hygiene	['(perform water treatment – 69.3)', '(carry waste water treatment – 59.5)', '(perform water treatment procedure – 58.4)']
water quality	technics of water quality	['(water consumption – 58.7)', '(manage air quality – 57.8)', '(advise soil water protection – 56.6)']

Green term	Green mention	Top 3 associated ESCO green terms
water recycling	recycling of maintenance water	['(water reuse – 74.1)', '(educate recycling regulation – 50.0)', '(maintain recycling record – 45.8)']
water supply	design of water supply infrastructure	['(water policy – 61.3)', '(advise soil water protection – 60.2)', '(water reuse – 59.7)']
water treatment plant	maintenance of a water treatment plant	['(maintain water treatment equipment – 74.7)', '(perform water treatment – 73.7)', '(perform water treatment procedure – 73.6)']
water treatment plant	understanding of water treatment plant	['(maintain water treatment equipment – 74.7)', '(perform water treatment – 73.7)', '(perform water treatment procedure – 73.6)']
wind energy	undertaking a wind energy project	['(renewable energy technology – 66.0)', '(offshore renewable energy technology – 62.2)', '(solar energy – 56.5)']
wind turbine	design of wind turbine	['(install onshore wind energy system – 69.3)', '(operate steam turbine – 65.0)', '(design offshore energy system – 59.1)']

Annex 4.
Source information for fields of study

Table 9. **Sources of information used in the classification of fields of study**

Official (ISCED)	https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-fields-of-education-and-training-2013-detailed-field-descriptions-2015-en.pdf
	https://uis.unesco.org/en/topic/international-standard-classification-education-isced
Official (NSI)	Spain (INE) www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736177034&menu=ultiDatos&idp=1254735976614
	Germany (DFG) www.dfg.de/download/pdf/dfg_im_profil/zahlen_fakten/programm_evaluation/faechersystematik_stabu_en.pdf
	Czechia(CSO) https://apl2.czso.cz/iSMS/en/klasstru.jsp?kodcis=80091&cisjaz=203
	United Kingdom (ONS) https://uis.unesco.org/en/topic/international-standard-classification-education-isced
	France www.paysdelaloire.fr/sites/default/files/typo3/tx_oxcsnewsfiles/Nomenclature_des_niveaux_de_formation.pdf
	France https://www.insee.fr/en/metadonnees/definition/c1746
	Hungary www.oktatas.hu/
	Belgium (Statbel) https://statbel.fgov.be/nl/open-data/code-isced-f-2013-4-cijfers
	Sweden www.scb.se/en/
	Poland https://education.org.pl/ https://nawa.gov.pl/images/users/623/Education_System_Poland_NAWA---2020-08-14_3.pdf
Other	Wikipedia https://en.wikipedia.org/wiki/International_Standard_Classification_of_Education
	(in different languages)
	https://en.wikipedia.org/wiki/National_Classification_of_Levels_of_Training

Delivering evidence from online job advertisements

Tapping into 10 years of experience

Online job advertisement (OJA) data have emerged as a crucial resource for identifying employers' skills needs. In collaboration with Eurostat, Cedefop has developed a comprehensive system for analysing OJA data, enhancing skills intelligence and generating experimental statistics on skills. This publication traces a decade of progress, from basic data collection to detailed analyses of occupations and skills. It provides a structured overview of OJA-based labour market intelligence, beginning with a framework for understanding OJA data and an in-depth look at the data production system. It further describes refined methodologies for skill and occupation classification, emphasising digital skills and the green transition. The richness of information provided by OJAs is presented even beyond skills. OJAs' ability to provide information on education fields and qualifications is also examined. Finally, the report consolidates insights, demonstrating how OJA-based intelligence supports labour market analysis and evidence-based policymaking, ensuring a data-driven approach to workforce development.

5610 EN - TI-01-25-049-EN-N - doi:10.2801/5070484

CEDEFOP | European Centre for the Development of Vocational Training

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

Publications Office of the European Union