International Labour Organization

ILO brief

Methodological Brief

January 2025

Developing a New Method to Uncover Skills Trends in Emerging Economies Using Online Data and NLP techniques*

Willian Adamczyk†, Simon Boehmer†, Isaure Delaporte†, Verónica Escudero†, Hannah Liepmann†

Key points

- This methodological brief describes an innovative approach that exploits online big data from job vacancies and applicants' profiles and utilizes natural language processing (NLP) to extract information on skills.
- The approach is based on a taxonomy comprising 15 unique skills subcategories across the broader cognitive, socio-emotional, and manual skills categories. It focuses on skills that are transferable rather than occupation-specific.
- So far tested with data on Uruguay, Brazil, the Russian Federation, and South Africa, the method is applicable to countries of all income levels and allows detecting country-specific skills trends. It captures a wide range of sectors and occupations, including those requiring manual labour.
- It can be used to offer novel insights into skills demand, supply and mismatch, as well as into the relationship between skills and aspects of job quality and/or job transitions.

Introduction

As global labour markets undergo rapid transformation, understanding the dynamics of skills is crucial for informed policy-making and economic development. In Western contexts, considerable scholarly efforts have been directed towards employing skills classifications to discern overarching skill trends and examine how skills dynamics influence wages and employment (see Autor, Levy, and Murnane 2003; Acemoglu and Autor 2011; Deming and Kahn 2018; Atalay et al. 2020; Hanushek et al. 2017). These studies have yielded critical insights into the influence of skill formation on workforce productivity, income distribution, and economic resilience. However, in contrast to the substantial focus on Europe and especially the United States, the exploration of similar inquiries remains relatively underdeveloped in emerging economies.

The existing literature draws upon diverse sources of information on skills. One prominent approach has been to use detailed occupational classification systems such as the Occupational Information Network (O-NET) of the United States and the European Skills, Competences, Qualifications and Occupations (ESCO) classification. However, the transferability of these occupational classifications to different national contexts is challenging,

* We would like to thank Evgeny Gushchin, Elvire Jégu and Franziska Riepl who provided research support for this brief.

† Skills, Active Labour Market Policies and Policy Evaluation Team at the Research Department of the ILO.

given that the skills composition within occupations differs significantly across countries. In addition, O-NET and ESCO represent extremely granular ontologies (i.e., comprehensive listings of relevant skills) that require conceptual aggregation before they can be used in research. Another approach utilizes surveys designed to measure skills, such as the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) and the World Bank's Skills Toward Employment and Productivity (STEP) surveys (OECD 2019; World Bank 2014). These initiatives offer valuable insights. However, their limited coverage in terms of countries and years—often a single year per country—constrains their applicability for achieving a comprehensive global perspective. Given the dynamic nature of tasks and skills in response to evolving labour markets, there is a need to longitudinally assess skills within and across occupations as well as in various country contexts.

This methodological brief presents an innovative approach which was first developed in Escudero, Liepmann, and Podjanin (2024) and applied on Uruguayan data, and then further refined and extended to additional countries—Brazil, the Russian Federation, and South Africa—for the forthcoming 2026 World Employment and Social Outlook (WESO) Report on Lifelong Learning and Skills Dynamics. This method leverages big online data and advanced natural language processing (NLP) techniques to reveal skills trends in low- and middle-income countries. The purpose of this brief is hence to provide a detailed description of the methodological framework, including data processing and analysis steps. It is meant not only to enable the replication of analyses featured in the WESO Thematic report and accompanying research projects, but also to support adaptation of the methodology for further research on skills trends in countries of different income levels. This brief thus aims to serve as a practical guide for researchers and practitioners seeking to apply these techniques in their own studies.

The use of online data on labour markets has in recent years increased substantially (see the discussion in Fabo and Kureková 2022). The methodology presented in this brief offers several distinct advantages. First, it leverages the nature and granularity of the data. By extracting detailed information from job vacancies and applicants' profiles, this approach addresses significant data gaps, enabling country-specific analyses without assuming occupational skills are similar across countries. It also allows to measure skills over time as they are presumably evolving in response to labour market transformations. Second, the proposed taxonomy is particularly adaptable to the realities of low- and middle-income countries. Unlike prior skills taxonomies for online vacancy data, which primarily focus on professional jobs in high-income countries (see, for example, the seminal study of Deming and Kahn 2018), this taxonomy incorporates manual skills that are relevant across countries, but even more so in low- and middle-income economies. It also expands the conceptual foundation of socio-emotional skills. Third, the taxonomy prioritizes transferable skills—those applicable across occupations—over occupation-specific skills, while highlighting numerous technical skills, making it particularly useful for understanding broad labour market trends.

In summary, this methodology provides a robust framework for analysing labour market dynamics. This in turn facilitates targeted interventions and policy formulation to enhance workforce development and economic growth.

The remainder of this brief presents the key elements of the methodology applied to create skills variables. It also discusses how, in a similar way, open-text descriptions can be used to extract information and create an ISCO-08 occupation variable. The methodology has already been applied to answer pertinent skills-related questions in a number of studies within the ILO (see Escudero and Riepl 2024 using South African and Uruguayan data; De Marzo, Mathew, and Sbardella 2023 using Indian data; and Escudero, Liepmann, and Vergara 2024 using Uruguayan data).

▶ Methodology to create skills variables

The methodology encompasses skills related to both job-specific tasks and personal attributes. This inclusive approach ensures a comprehensive representation of skills-related dynamics across diverse labour markets, covering the skills sought by employers in job postings and those highlighted by workers in their online profiles. In the following, the basic building blocks of the original taxonomy—developed in Escudero, Liepmann, and Podjanin (2024)—are presented along with the refinements that were implemented in further work.

Taxonomy

In the taxonomy, skills are grouped into three broad categories, namely cognitive, socio-emotional and manual skills. Within each category, further distinct skills are identified to arrive to a classification comprising 15 subcategories (14 in the original concept), making it a nuanced yet compact taxonomy.

The taxonomy is built upon existing literature from labour economics and psychology, but has been expanded to make it adaptable to individual country-contexts, with a particular focus on emerging and developing countries and applications in online data. Initially, the skills categorization was developed based on established taxonomies designed for classifying skills in online data within the United States (U.S.), particularly Deming and Kahn (2018).[1] The taxonomy was then expanded to include manual skills, which are typically omitted in U.S.-centred analyses of online data. The taxonomy includes three skills subcategories for manual skills (i.e. finger dexterity, hand-foot-eye coordination, and physical skills) to facilitate a more comprehensive analysis of online data beyond individuals with high formal qualifications. Moreover, the conceptual foundations of cognitive and socio-emotional skills were broadened by integrating additional keywords that capture shifts in terminology over time while still representing the same skill (Deming and Noray 2020).

To achieve this, a dictionary of keywords and expressions referring to each skill subcategory was developed. These keywords and expressions were drawn from various seminal studies, some of which do not rely on online data sources (Almlund et al. 2011; Atalay et al. 2020; Deming and Noray 2020; Heckman and Kautz 2012; Hershbein and Kahn 2018; Kureková et al. 2016; Spitz-Oener 2006), as well as the pilot version of O-NET Uruguay.[2]

As mentioned, further improvements were made to the initial version of the taxonomy. The changes include:

1. Splitting the cognitive skills subcategory into core cognitive and sophisticated cognitive skills. The updated version of the taxonomy thus includes 15 subcategories of skills.
2. Certain skills subcategories have also been refined: i) the former project management skills subcategory has been renamed to project and process management skills and now comprises additional types of skills; ii) as a result, the people management skills subcategory has been refined and now refers to skills only specific to the management of people.
3. Lastly, the list of keywords and expressions has been updated following a new round of context revision. For instance, to account for modern technologies, programming language and software, additional terms were retrieved from topics in Stack Overflow and Github, the two most popular platforms amongst software developers, and were added to the list of keywords. Similarly, expressions related to machine learning and artificial intelligence were added to the list of keywords using relevant tagged questions in Stack Overflow.

Table 1 presents an overview of the skills categories and subcategories, along with their definitions and sources they were derived from. Additionally, Table A1 in the Appendix lists the most important identifying keywords and details the changes made compared to the original version presented in Escudero, Liepmann, and Podjanin (2024).[3] For more details, researchers and practitioners seeking to replicate the method are encouraged to contact the ILO through the contact details provided at the end of this brief.

[1] Other sources used include Deming and Noray (2020); Heckman and Kautz (2012) and Kureková et al. (2016).

[2] O-NET Uruguay is an occupational classification system which describes the tasks and skills associated with a specific occupation in the Uruguayan context, similar to the US O-NET. When the methodology of this brief was developed, the O-NET Uruguay pilot project captured 22 selected occupations only (see Ministerio de Trabajo y Seguridad Social 2020; Velardez 2021).

[3] The complete dictionary for each language will be made publicly available as part of the work undertaken for the forthcoming 2026 World Employment and Social Outlook (WESO) Report on Lifelong Learning and Skills Dynamics.

▶ Table 1: Categorization of skills, keywords, and sources

Subcategory	Definition	Sources
	Cognitive skills	
Cognitive skills (core)	Skills needed to perform tasks that require analysis and calculation, problem-solving, intuition, and flexibility.	**DK (2018)**; ALM (2003); DN (2020); S-O (2006)
Cognitive skills (sophisticated)	Skills needed to perform more sophisticated tasks that require analysis, modelling, and creativity.	**DK (2018)**; ALM (2003); DN (2020); S-O (2006)
General computer skills	These six subcategories relate closely to the cognitive skills described above. They correspond to skills that are needed in specific areas of work. They are listed as separate subcategories because they are often specifically mentioned in job adverts and in applicants' work experiences. They are mostly geared towards white-collar jobs, in line with the aim of the work by Deming and Kahn (2018).	APST (2020); O-NET Uruguay; **DK (2018); DN (2020)**
Software skills and technical support		
Machine learning and AI		
Financial skills		
Writing skills		
Project and process management		
	Socio-emotional skills	
Character skills	Character skills include three of the five categories of the five-factor personality model commonly used in the psychological literature (McCrae and Costa 2008). It includes conscientiousness, openness to experience, and emotional stability. In addition, this subcategory includes dimensions such as being relaxed, independent, self-confident and the degree of vulnerability to stress.	**DK (2018);** DN (2020); KBHT (2016); HK (2012)
Social skills	Social skills include those character traits from the five-factor personality model that are less related to one's personal attributes and related more to how one interacts with other people, specifically agreeableness and extraversion. Other keywords that relate to the general ability to have personal interactions, such as working in teams or holding presentations, are included as well.	**DK (2018);** DN (2020); KBHT (2016); HK (2012); S-O (2006); APST (2020); O-NET Uruguay
People management skills	Lastly, two subcategories are added that refer to specific abilities within the broader realm of social interactions, and which are often listed as particular requirements in job adverts and applications.	**DK (2018);** DN (2020);
Customer service skills		ALM (2003); S-O (2006)
	Manual skills	
Finger-dexterity skills	This category focuses on manual skills that are usually classified as "routine" by ALM (2003) and which are common in machine operation and the production or handling of goods. Examples include picking and sorting in agriculture or working in an assembly line.	**ALM (2003);** S-O (2006); APST (2020); O-NET Uruguay
Hand-foot-eye coordination skills	These manual skills are usually understood as "non-routine" by ALM (2003). They are more commonly used in services-related occupations and include working in changing environments that necessitate adaptation. This category encompasses for example driving cars or repairing and cleaning items.	**ALM (2003);** S-O (2006); PST (2020); O-NET Uruguay
Physical skills	This subcategory focuses on more innate bodily characteristics, such as physical strength, endurance, the ability to lift heavy objects or work while standing or walking.	**O-NET Uruguay**

Source: Table 1 of Escudero, Liepmann, and Podjanin (2024), where additional information on the conceptual background and concrete keywords used in the taxonomy can be found. The latest list of keywords and expressions can be found in Table A1 of the Appendix.

Notes: ALM (2003) stands for Autor et al. (2003), APST (2020) for Atalay et al. (2020), DK (2018) for Deming and Kahn (2018), DN (2020) for Deming and Noray (2020), HK (2018) for Hershbein and Kahn (2018), HK (2012) for Heckman and Kautz (2012), KBHT (2016) for Kureková et al. (2016) and S-O (2006) for Spitz-Oener (2006). To keep pace with changing market dynamics, the skills taxonomy and associated dictionaries should be regularly updated and revised.

Implementation

This taxonomy is then applied to the online vacancy and applicants' data combining rule-based classification—derived from the taxonomy—with natural language processing (NLP) techniques. These techniques allow to systematically access and review unstructured text information contained in big data. More specifically, open-text descriptions of vacancies are pre-processed to fit the structured format of the skills taxonomy and its list of keywords. Skills are identified based on the specific dictionary of keywords and expressions associated with each of the 15 skills subcategories in the taxonomy. Box 1 provides an overview of the data used so far.

▶ Box 1. Data

The methodology was applied and adapted to four country contexts: Uruguay, Brazil, the Russian Federation, and South Africa. The data used comes from the following sources:

- BuscoJobs, a private job-search portal where firms can post vacancies for a small fee and applicants can create profiles and apply to the vacancies. It provides detailed information on job vacancies posted by firms, applicants searching for jobs, and on the applications made by job seekers to those vacancies in Uruguay and Brazil. The data we use span years 2010 through 2023.
- Adzuna, a job aggregator which collects, standardizes and re-posts vacancies published on the internet. It provides detailed job advert data in the Russian Federation and South Africa. For both countries, we use data covering the period from April 2016 through December 2021.

Although this methodology has been implemented and tested specifically in the context of these four countries, the growing availability of online data presents an opportunity for applying the methods elsewhere. As such, the methodology holds significant potential for generating insights into country-specific labour dynamics across a broader array of global contexts. One such example is the study by De Marzo, Mathew, and Sbardella (2023), which uses vacancy data to investigate the link between skills demand and firms' productivity and innovation in India.

Although some skills subcategories are closely linked, the keywords and expressions used to characterize them are distinct and mutually exclusive, which allows for the unique identification of skills in the data. When keywords overlap in different categories, an exception rule was added to avoid partial matches. Such cases are for instance: "design" which is included in core cognitive skills and "design site" which is in software and technical support skills; or "repair" in hand-foot-eye coordination skills and "computer" in computer general skills, which do not match "repair computer" in software and technical support skills; or lastly, "team" in social skills, which is treated separately from "lead team" or "team management" in people management skills.

To adapt the dictionary to country-specific languages and contexts, keywords were translated while considering local expressions. Some modifications were made to the original keywords to avoid ambiguity or miscontextualization after they had been translated. To ensure correct usage, rounds of manual inspection were conducted by the authors of this brief to verify whether the suggested word was suitable in at least 75 per cent of descriptions in 40 randomly selected observations. If it was found not to be suitable, whenever feasible another synonym or compound expression was chosen instead of dropping altogether the directly translated keyword.

For instance, the keyword "running", as listed in the taxonomy under physical skills, can be translated as "correr" in Spanish and Portuguese dictionaries, which accurately reflects the physical context. In English, however, the word "running" can refer to non-physical tasks as well such as "running a project". Therefore, this keyword was omitted from the English dictionary. Synonyms like "run(ning) fast" or "athletic (run)" were considered non-useful as they did not appear frequently in job postings or were used in incorrect contexts. Still, the physical aspects of "running" were conveyed through terms like "walking", "strolling", "hiking", or "marching".

The dictionaries were developed specifically for each country as follows:

• **Uruguay**: This dictionary was developed and implemented first, based on the translation to Spanish of the English dictionary that was elicited from the literature review in Escudero, Liepmann, and Podjanin (2024). After translation, terms that consisted of two words were also added in reverse order and synonyms were added based on the Spanish version of the synonym website www.wordreference.com. The added synonyms were

manually reviewed to discard out-of-context or overlapping terms. This yielded first a set of 275 keywords and expressions[4] (without synonyms), which increased to a final set of 669 keywords and expressions after synonyms had been included.

• **Brazil**: the original English dictionary was translated to Portuguese, taking into account the adjustments that were made for the Spanish version. Synonyms from the website www.sinonimos.com.br were added, yielding first a total of 306 keywords and expressions, and finally 1,339 keywords and expressions including synonyms.

• **The Russian Federation**: The English dictionary was translated into Russian taking into account potential differences in context.[5] In an additional step, software and company names were also added in the Latin alphabet, and selected keywords with ambiguous meanings were modified or dropped. Synonymys were scraped from https://synonymonline.ru/ and refined manually. The final dictionary contains 303 keywords and expressions and 1,951 such terms when including synonyms.

• **South Africa**: This dictionary was built directly from the English keywords derived from the literature. Synonyms were added by performing an automated web scraping of the synonym website www.wordreference.com. This procedure yielded a final set of 284 keywords and expressions or 1,686 with the inclusion of synonyms.

Overall, web scraping procedures return a different number of synonyms for each language, owing to its specific linguistic characteristics. The obtained synonyms were reviewed manually to confirm their relevance and to ensure comparability across countries. Nevertheless, the different number of synonyms does not substantially affect the number of matched skills for each country.

Creation of the skills variables

The final step involves creating the skills variables by leveraging the unstructured text data found in vacancy descriptions posted by firms and the job spells listed in applicants' profiles. The open-text descriptions offer a viable approach for creating skills variables, as they contain detailed information on skills for all vacancies and a majority of applicants' job spells.

The open-text descriptions undergo a series of NLP pre-processing steps, including: (i) tokenization (i.e., splitting the texts into their individual words to allow for further processing), (ii) text normalization (e.g., converting to lowercase and removing accents, numbers, special characters, and words with fewer than two letters), (iii) stop words removal (i.e., dropping words that do not carry meaning for the exercise at hand), and (iv) lemmatization (i.e., reducing words to their common root to unify variations of the same concept, such as "communicate" and "communication", while accounting for gender, plural, and verb tense variations). These processes are applied to both the online data and the skills taxonomy categorizations, to facilitate the mapping of the two.

Finally, each skill subcategory is considered and coded as present if at least one of the keywords from the dictionary is identified in the text. Additionally, the frequency of keyword occurrences for each skill subcategory is calculated[6] and used to study the overall supply and demand for skills.

The NLP implementation largely follows the same process in all four countries, except when different sets of tools and algorithms were necessary to accommodate data size and language-specific algorithms:

• **Uruguay**: as the total number of observations was 164,864 vacancies and 1.5 million job spells, the data was processed locally using Python's Natural Language Toolkit (NLTK) to tokenize, normalize, drop stopwords and lemmatize words.

• **Brazil**: as the data comprises 42 million vacancies and 3.2 million applicants' job spells, PySpark and SparkNLP were required to tokenize, normalize, drop stopwords and lemmatize words.

• **The Russian Federation**: the data comprises 170 million observations for vacancies (there is no information on applicants). However, due to performance restrictions, a random sample of 10 per cent of the data was used. With 17 million, the resulting sample is large enough to detect all relevant patterns in the data. Similarly to the case of

[4] Keywords refer to single words, whereas expressions denote multiple words belonging together.

[5] The contributions of Evgeny Gushchin to the skills variable creation for the Russian Federation are gratefully acknowledged.

[6] While this frequency provides valuable insights, it does not fully capture the intensity with which a particular skill is used. This constitutes a caveat that warrants further exploration in future research.

Brazil, PySpark and SparkNLP were required to tokenize, normalize, drop stopwords and lemmatize words.

• **South Africa**: the data contains 6.2 million observations (again only of vacancies). This sample size required using the big data tools present in PySpark and SparkNLP.

Turning to the classification outcomes, applicants' descriptions of their past jobs tend to be shorter than vacancy postings for similar jobs, and thus capture less skills on average. Nevertheless, the method succeeds in capturing a significant number of skills for both applicants and vacancies (see Table 2 for descriptives statistics by country). Moreover, the number of assigned skills increases when considering the applicant level (i.e., aggregating information from a person's job spells) rather than the individual job spell level (Escudero, Liepmann, and Podjanin 2024).

▶ Table 2: Descriptive statistics

	Uruguay	Brazil	Russian Federation	South Africa
Vacancies				
Share of ads in which at least one skill was identified	137,773 (83.6%)	35,882,914 (85.3%)	13,513,673 (80.6%)	5,262,557 (84.9%)
Average number of skills	3.14	2.84	2.76	4.53
Average length of job ads	60.80	116.20	89.99	139.9
Applicants				
Share of spells with at least one skill	553,533 (36.2%)	1,212,651 (37.4%)		
Share of applicants with at least one skill	284,354 (31.9%)	538,755 (24.4%)		
Average number of skills	1.71	2.11		
Average length of job spell	17.13	41.90		

Source: Analysis based on BuscoJobs data for Uruguay and Brazil; and Adzuna data for the Russian Federation and South Africa.

▶ Methodology for ISCO-08 classification

The data obtained from online big data sources often does not contain an occupational variable. In the case of the raw data provided by BuscoJobs (Uruguay and Brazil), it only classifies vacancies and applicants' job spells into the International Classification of Occupations (ISCO)-08 for a limited subsample. In the raw data obtained from Adzuna, for South Africa, job titles were standardized, but no formal ISCO assignment was carried out. In the case of the data for the Russian Federation, there had been no standardization effort nor ISCO assignment.

Yet, conducting analyses at the occupational level is crucial to better understand skills dynamics across and within occupations. A methodology similar to the one used for skills was applied to create ISCO-08 occupation variables for both vacancies and applicants' job spells. This process, outlined in Escudero, Liepmann, and Podjanin (2024) for Uruguay, forms the basis for the following summary. The creation of this variable involves several steps, which are discussed below[7]:

• **Uruguay:** for vacancies, textual information is gathered from four open-text fields: the job title, job description, required level of education, and the hierarchical level of the vacancies. For job spells in applicants' work histories, the same information is used, except from job titles, which are not available as a separate entry.

The data undergoes NLP processes similar to those applied in the creation of the skills variables. The resulting text is then categorized into ISCO-08 occupational categories through a three-step process. The first step which relies on a rule-based model matches job titles to keywords, which are selected from the most frequently used words and phrases in both vacancies and applicants' job spells that

[7] The contributions of Giordano de Marzo for creating occupational variables in the Brazilian and Russian data and of Fidel Bennett, Sergio Herrera and Javiera Lobos for creating these variables in the Uruguayan data are gratefully acknowledged.

were already classified by BuscoJobs into ISCO-08 occupations. This exercise, which included in some cases a manual re-classification, establishes a set of rules to classify the remaining job titles into 2-digit ISCO occupational categories.

As a second step, educational level information is used to distinguish between ISCO skill levels 2 and 3, individuals within the same field who possess higher education (level 2) or any other educational levels (level 3). Similarly, information about the hierarchical level is used to identify managers and directors, assigning them in level 1 of the ISCO classification.

To enhance the performance of the procedure, a third step is conducted, introducing a machine learning algorithm (in the form of a predictive model) to assign vacancies and job spells that remained unclassified or for which the original BuscoJobs assignment significantly differs from that of the algorithm. This process occurs in two steps. Initially, the model is trained using the already classified observations to assign 1-digit ISCO codes. Subsequently, additional information from applicants and vacancies is incorporated into a second prediction model to further refine the assignment at the 2-digit level.

After various performance assessments, the Gradient Boosting algorithm was chosen to code 1-digit and 2-digit occupations in the vacancy data, and the Random Forest for the applicants' data. Overall, all vacancies could be classified at the 1-digit level, with around 95 per cent also being classified at the 2-digit level. For applicants, all job spells with a text description were classified at the 1-digit level, and around 98 per cent were also classified at the 2-digit level. For additional details regarding this procedure and the underlying keywords, please refer to Appendix B of Escudero, Liepmann, and Podjanin (2024).

In a subsequent version of the dataset, which included additional data for the years 2021 to 2023, an additional step was introduced. This fourth step utilized the previously classified data as a training sample to train the same machine learning models. The aim of this step was to extrapolate the existing classifications to the job adverts and applicants' job spells from 2021 to 2023, where no ISCO classification was provided by BuscoJobs.

It is worth noting that the creation of the ISCO variable involved slightly different steps across countries, depending on the specific data available:

• **Brazil and the Russian Federation:** The 2-digit ISCO-08 variables were generated using a novel experimental approach leveraging Generative Artificial Intelligence (AI) and "Bidirectional Encoder Representations from Transformers (BERT)"-embeddings. The process involved three main steps. First, ChatGPT 4.0 was instructed to generate standardized job adverts for all ISCO 2-digit codes, providing a robust training sample for the machine learning algorithms. In a second step, job descriptions were cleaned through an NLP process, and BERT embeddings were applied to capture the semantic meaning of sentences describing job requirements and tasks. In a third step, job adverts for Brazil and the Russian Federation, as well as applicants' job spells for Brazil, were classified into 2-digit ISCO-08 categories using Random Forest classifiers.

• **South Africa:** the 2-digit ISCO-08 variable was built through a combination of automatic classification using the normalized job titles and a manual revision based on ISCO's documentation. Initially, 4,553 distinct job titles were extracted from the raw data and automatically classified using the R package labouR, designed for English-language data. A manual revision was then conducted, where 932 occupations were manually recoded to ensure alignment with ISCO-08 standards. This revision considered: i) whether the occupation was mentioned in ISCO-08 documentation; ii) its similarity to documented occupations; and iii) the expected skill level for each job following the ISCO documentation.

▶ Conclusions

This methodological brief described in detail an innovative approach that utilizes big data from online job boards and job aggregators and applies NLP processes to extract information on skills. Originally developed in Escudero, Liepmann, and Podjanin (2024), the methodology has since been refined and expanded to other countries for the forthcoming 2026 World Employment and Social Outlook (WESO) Report on Lifelong Learning and Skills Dynamics.

The methodology is highly adaptable and can be applied across various country contexts. By avoiding the assumption that the task composition of occupations is uniform across countries or constant over time, it enables a more nuanced analysis of the unique characteristics of different labour markets and their corresponding policy needs.

Furthermore, it allows for the identification of emerging skills, such as green or digital skills, which are critical in

times of global transformations. While the aggregation of skills into subcategories is well-suited for research analyses, the methodology also permits disaggregation into more detailed groups. This granularity supports assessments of specific career pathways by identifying the skills individuals possess and how these skills complement others, providing valuable insights for skills development and targeted policy interventions.

However, it is important to acknowledge certain caveats when using this data. While it leverages the granularity of information not typically available from traditional sources, its representativity may vary depending on factors such as the platform used, the population accessing online job boards, and the completeness of the data for certain industries or regions. Therefore, findings based on this methodology should be interpreted with careful consideration of these limitations.

The methodology supports a broad range of analyses on skills dynamics, as illustrated by the studies conducted so far using these data and methodology (see Escudero, Liepmann, and Podjanin 2024; Escudero and Riepl 2024; De Marzo, Mathew, and Sbardella 2023; Escudero, Liepmann, and Vergara 2024). As the upcoming WESO report will further demonstrate, it facilitates analyses both within and across occupations, addressing a critical gap in the literature, which often emphasizes skills comparisons solely across occupations. By providing precise and nuanced insights, this methodology holds significant potential to inform the design of effective and sustainable labour market policies.

References

Acemoglu, Daron, and David Autor. 2011. 'Skills, Tasks and Technologies: Implications for Employment and Earnings'. In *The Handbook of Labor Economics*, O. Ashenfelter and D. Card (eds), 4:1043–1171. Amsterdam: Elsevier.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz. 2011. 'Chapter 1 - Personality Psychology and Economics'. In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 4:1–181. Handbook of The Economics of Education. Elsevier.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2020. 'The Evolution of Work in the United States'. *American Economic Journal: Applied Economics* 12 (2): 1–34.

Autor, David, Frank Levy, and Richard Murnane. 2003. 'The Skill Content of Recent Technological Change: An Empirical Exploration'. *The Quarterly Journal of Economics* 118 (4): 1279–1333.

De Marzo, Giordano, Nanditha Mathew, and Angelica Sbardella. 2023. 'Who Creates Jobs with Broad Skillsets? The Crucial Role of Firms'. Working Paper 94. ILO Working Paper.

Deming, David, and Lisa Kahn. 2018. 'Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals'. *Journal of Labor Economics* 36 (S1): S337–69.

Deming, David, and Kadeem Noray. 2020. 'Earnings Dynamics, Changing Job Skills, and STEM Careers'. *Quarterly Journal of Economics* 135 (8): 1965–2005.

Escudero, Verónica, Hannah Liepmann, and Ana Podjanin. 2024. 'Using Online Vacancy and Job Applicants' Data to Study Skills Dynamics'. *Research in Labor Economics* 52B:35–99.

Escudero, Verónica, Hannah Liepmann, and Damian Vergara. 2024. 'Directed Search, Wages, and Non-Wage Amenities: Evidence from an Online Job Board'. ILO Working Paper.

Escudero, Veronica, and Franziska Riepl. 2024. 'Revealing New Skills Trends in Emerging Economies: The Power of Online Data and NLP Techniques | International Labour Organization'. 8 August 2024. https://www.ilo.org/publications/revealing-new-skills-trends-emerging-economies-power-online-data-and-nlp.

Fabo, Brian, and Lucia M. Kureková. 2022. 'Methodological Issues Related to the Use of Online Labour Market Data'. ILO Working Paper 68. Geneva: International Labour Organization.

Hanushek, Eric A., Guido Schwerdt, Ludger Woessmann, and Lei Zhang. 2017. 'General Education, Vocational Education, and Labor-Market Outcomes over the Lifecycle'. *Journal of Human Resources* 52 (1): 48–87.

Heckman, James, and Tim Kautz. 2012. 'Hard Evidence on Soft Skills'. *Labour Economics* 19 (4): 451–64.

Hershbein, Brad, and Lisa B. Kahn. 2018. 'Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings'. *American Economic Review* 108 (7): 1737–72.

Kureková, Lucia Mýtna, Miroslav Beblavý, Corina Haita, and Anna-Elisabeth Thum. 2016. 'Employers' Skill Preferences across Europe: Between Cognitive and Non-Cognitive Skills'. *Journal of Education and Work* 29 (6): 662–87.

Ministerio de Trabajo y Seguridad Social. 2020. 'Análisis Primario de Resultados de La Primera Ola de Relevamiento Del Perfil de Ocupaciones - O*Net'. Unpublished Report. Montevideo.

OECD. 2019. *Skills Matter: Additional Results from the Survey of Adult Skills*. Paris: OECD Publishing. https://doi.org/10.1787/1f029d8f-en.

Spitz-Oener, Alexandra. 2006. 'Technical Change, Job Tasks, and Rising Educational Demands: Looking Outside the Wage Structure'. *Journal of Labor Economics* 24 (2): 235–70.

Velardez, Miguel Omar. 2021. 'Análisis de distancias ocupacionales y familias de ocupaciones en el Uruguay'. Documento de Proyectos LC/TS.2021/36. Desarrollo Económico. CEPAL.

World Bank. 2014. 'STEP Skills Measurement Snapshot 2014'. https://www.worldbank.org/content/dam/Worldbank/Feature%20Story/Education/STEP%20Snapshot%202014_Revised_June%2020%202014%20%28final%29.pdf.

Appendix

▶ **Table A.1: Categorization of skills, keywords, and sources**

Category	Selected keywords/expressions	Key changes since Escudero et al. (2024)
	Cognitive skills	
Cognitive skills (core)	problem solving, analytical, critical thinking, adaptability, direction, planning, decision making, interpreting rule, processing information, design, sketch, idea generation, calculation, bookkeeping, correcting, measurement, memory	Cognitive skills were split into core and sophisticated
Cognitive skills (sophisticated)	research, science, evaluate, devise rule, statistics, math(ematics), data analysis, data engineering, data modelling, data visualization, data mining, predictive model	
General computer skills	computer, internet, spreadsheet, software, MS Windows, MS Excel, MS PowerPoint, MS Word, MS Office, Outlook,	The word "software" is being considered in Computer (general) skills to avoid overlap
Software (specific) skills and technical support	programming, Java, SQL, Python, Javascript, HTML, PHP, Typescript, Swift, Kotlin, Scala, Ruby, computer installation, computer repair, computer maintenance, computer troubleshooting, web development, site design	Expressions for new technologies were added and some lemmas were corrected
Machine Learning and Artificial Intelligence	artificial intelligence, automation, machine learning, reinforcement learning, (un)supervised learning, deep learning, natural language processing (NLP), computer vision decision trees, random forest, cluster analysis, neural networks, Bayesian networks, convolutional neural network (CNN), support vector machines (SVM), TensorFlow, MapReduce, Splunk, Keras, PyTorch, PySpark, Apache Spark, Apache Hadoop	
Financial skills	budgeting, accounting, finance, cost Softland, ERP, SAP, Xubio, Cloudboooks, Nubox, Bloomberg, Anfix	
Writing skills	writing, editing, reports, proposals	

Project and process management skills	project management, commercial management, process management, purchasing management, stock management, operation management, product management, supply management	Initially considering only “Project management”, the scope of the category was enlarged to also include Process management. The following words were added: “commercial management“, “process management“, “purchasing management“, “stock management“, “operation management“, “product management“, “supply management“
Socio-emotional skills		
Character skills (conscientious-ness, emotional stability and openness to experience)	organized, detail oriented, multitasking, time management, deadlines, energetic, self-starter, initiative, self-motivated, competent, achieving, hardworking, reliable, punctual, stress resistant, creative, independent	
Social skills (including agreeableness and extraversion)	communication, teamwork, collaboration, negotiation, presentation, team, persuasion, listening, flexibility, empathy, assertiveness, advice, entertain, lobby, teaching, interact with others, verbal abilities	
People management skills	supervisory, leadership, lead team, people management, team management, coordinate team, mentoring, staff management, staff supervision, staff development, performance management, personnel management, human resource management	Words “management” and “people” were dropped and added: “lead team“, “people management“, “team management“, “coordinate team“, “staff management“, “human resource management“
Customer service skills	customer, sales, client, patient advertise, sell, buy, purchase	Some lemmas were corrected in Spanish*.
Manual skills		
Finger-dexterity skills	picking, sorting, repetitive assembly, mixing ingredients, baking, sewing, trimming, decorating, tabulating machine, control machine, packing, equipment, welding	Word “product” was dropped, and other words were added: “equipment”, “welding”.
Hand-foot-eye coordination skills	attending cattle, attending animals, driving, transporting, piloting, pruning, gymnastics, sports, reaction time, fine manipulations, accommodate, renovate, repair, restore, serving, cleaning, replenish stock	Word added: “replenish stock”
Physical skills	resistance, carrying loads, unload, walking, physical strength, physical conditioning, physical endurance	Words were added: “unload”, “physical strength”, “physical conditioning”, “physical endurance”, “physically demanding”, “physically fit”.
Notes: See Table 1 and the main text for details. * Due to a particularity in Python's SpaCy lemmatizer for Spanish, alternative versions of lemmas had to be included in the dictionary for certain keywords. For instance, variations such as "cliente"/"client," "venta"/"ventar," and "soldadura"/"soldaduro" were considered to ensure comprehensive coverage and accuracy.		

Contact details
Verónica Escudero (escudero@ilo.org)
Isaure Delaporte (delaporte@ilo.org)

International Labour Organization
Route des Morillons 4
CH-1211 Geneva 22
Switzerland

T: +41 22 799 7239
E: @ilo.org

DOI: https://doi.org/10.54394/HQQX3200

© International Labour Organization 2025