SkiLMeeT

# Dataset online platforms

Authors: Roy Peijen, Sadegh Shahmohammadi, Joost van Genabeek, Diane Confurius, Janika Thielecke

Reviewers: Suhendan Adiguzel van Zoelen, Mairi Bowdler

Grant agreement no. 101132581

Deliverable: D2.2 Dataset online platforms

Due date: 31.03.2025

Funded by
the European Union

## Document control sheet

| | |
|---|---|
| **Project Number:** | **101132581** |
| **Project Acronym:** | SkiLMeeT |
| **Work-Package:** | WP2 |
| **Last Version:** | 26.03.2025 |
| **Issue Date:** | 31.03.2025 |

## Classification

| | |
|---|---|
| Draft | |
| Final | x |
| Confidential | |
| Restricted | |
| Public | x |

## Legal notice

*This project has received funding from the European Union's Horizon Europe Research and Innovation programme under grant agreement No. 101132581. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.*

## Table of contents

**1. Data collection .... 5**

1.1. Revelio Labs .... 6

1.1.1. Data Description .... 6

1.2. LinkedIn data .... 7

1.2.1. Method .... 7

1.2.2. Added value .... 8

1.2.3. Limitations .... 10

1.3. Social media campaigns to collect resumes .... 11

**2. References .... 11**

## Management Summary

SkiLMeeT research project, which investigates the evolving demand for skills in the context of the digital and green transitions within European economies, aims to identify skill shortages, their drivers, and propose innovative solutions to address these gaps.

This document provides an overview of the big data collected from online platforms to examine the supply and demand for skills across various EU countries. Data was sourced from Revelio Labs, LinkedIn, and social media campaigns to obtain insights on skills and job transitions. Revelio Labs provided a comprehensive dataset comprising over 1 billion professional profiles, encompassing detailed information on individual skills, education, and job transitions. LinkedIn's aggregated data offered valuable perspectives on career transitions, skill similarities, and transition popularity. Despite considerable investment, social media campaigns resulted in only 22 valid CVs.

**Keywords:** LinkedIn, Revelio Labs, big data, skill supply, skill demand.

# 1. Data collection

We collected data from several sources for this deliverable: (1) Revelio Labs data, (2) LinkedIn data, and (3) Social media campaigns to collect resumes. Table 1 represents a brief overview of the main information obtained. For each data source, we have outlined the methodology used to collect the data, the calculation of specific metrics, and the added value and limitations associated with each source.

Table 1. Brief overview of the data

| Source | Data type | Information | Added value | Limitations |
|---|---|---|---|---|
| Revelio Labs | Web scraped (LinkedIn data etc.) | • Individual profession profiles<br>• Company information<br>• Individual Skills<br>• Individual Education | • Rich individual information<br>• Data regarding job transitions | • Data is manually entered by the users (can lead to data incompleteness)<br>• Users may not be representative of the entire workforce |
| LinkedIn | Aggregated data | • Skill Similarity<br>• Popularity | • Self-reported skills, worker-defined abilities<br>• Skill connections, job/industry alignment<br>• LinkedIn data provides an opportunity to critically evaluate its own skill similarity metrics by comparing them with methodologies used in this study | • *Aggregation limitation:* Obscures individual variations, loss of granular insights, overlooks nuanced skill sets and career transitions<br>• Skills genome is updated continuously but lacks clear documentation. |
| Social media campaigns to collect resumes | Survey | | | • Because of luck of interest, the CV campaign is stopped in month 3. |

## 1.1. Revelio Labs

### 1.1.1. Data Description

The data from Revelio Labs contains information about individual professional profiles (from LinkedIn), the individual's educational history, the individual's skills information and company information. These reach data sources can offer valuable insights into the labour market trends.

The dataset contains more than 1 billion profiles, more than 20 million companies and more than 3 thousand skills and activities.

Each individual professional profile is described by a rich list of variables.

Table 2. Description of variables in user profile data set

| Variables for Individual User Profiles | Variable Definition |
|---|---|
| User_ID | User Identifier |
| F_prob, M_prob, sex_predicted | Probability of user being female/male |
| User_country | Original country of the user |
| Highest_degree | Education degree |
| numconections | Number of connections |
| rcid | Revelio Labs company ID |
| Startdate, enddate | Position start and end dates |
| Title_raw, title_translated, job_category, description | Raw job title; Job title translated to English; Main category of job (e.g. Engineer); Job description |
| Remote_suitability | Predicted remote suitability |
| Seniority | Seniority level (7 levels) |
| Salary, total_compensation | Modelled annual salary for the position in USD |
| Onet_code, onet_title | O*NET-SOC code and title (e.g. 11-1011.00 Chief Executives) |
| Naics_code, naics_description | Company's NAICS industry code and description |
| Ricks_k50, rics_k200, rics_k400 | Industry of employer with 50/200/400 discrete categories |
| Country, location_raw, region, state, metro_area | Country, region, state and metro area of the job position |

This extensive dataset of Individual User Profiles (286GB), which is stored across over 1k parquet files, was filtered by country and occupation to focus on specific regions and jobs of interest. The selected countries include the Netherlands, Germany, Luxembourg, Italy, Poland and Bulgaria. Two expert-curated lists were used to identify jobs of interest: one list with 138 green occupations and another with 53 digital occupations. These filtering criteria were applied to the entire dataset to ensure the selection of users belong to one of the countries and have worked in at least one job from each of the specified job lists.

Filtering the users for the selected countries and those who had at least one of the jobs in the list of green jobs resulted in a total of 4,974,467 unique users (68k for Luxembourg, 543k for Poland, 90k for Bulgaria, 1M for Netherlands, 1.9M for Germany and 1.3M for Italy).

Filtering the users for the selected countries and those who had at least one of the jobs in the list of digital jobs resulted in: a total of 3,542,036 unique users (55k for Luxembourg, 550k for Poland, 102k for Bulgaria, 809k for Netherlands, 1.2M for Germany and 793k for Italy).

Next, we will analyse the two generated subsets, focusing on job transitions across various countries for the selected occupations. This analysis offers several interesting avenues to explore, including understanding the primary drivers behind career switches, identifying which occupations are in higher demand, assessing skill gaps between jobs, evaluating geographic influences, examining factors contributing to employee retention and satisfaction, and more.

*Data has been obtained in accordance with all applicable laws and regulations regarding data protection, data security, and data privacy (“Applicable Data Protection Laws”).

## 1.2. LinkedIn data

### 1.2.1. Method

LinkedIn provided us with aggregated data on 2024 career transitions, using their own occupation classification (along with a crosswalk to ISCO-08 codes). This data includes the skill similarity metric and the popularity of each transition by country. The countries included are: Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom.

- *Skill Similarity:* This indicator quantifies how closely the skill sets of two occupations or industries align. LinkedIn calculates the distance between the skill sets of two entities to assess their similarity using the ‘skills genome’ methodology. For any entity (such as an occupation, country, or industry), the skills genome is a ranked list (vector) of the 50 most defining skills for that entity. The skills genome identifies these skills using a TF-IDF algorithm, which highlights the most representative skills for the entity while reducing the weight of common skills (e.g., Microsoft Word) that don't offer much distinguishing information. TF-IDF is a statistical method that measures how significant a skill is to a given entity by combining two components: (1) the term frequency (TF), which reflects how often a skill appears within the entity; and (2) the inverse document frequency (IDF), which measures how rare or common a skill is across a broader set of entities. A lower IDF indicates the skill is more common, while a higher IDF shows the skill is more unique. Thus, a skill that is widespread across LinkedIn profiles will have a low IDF, while a skill exclusive to certain entities will have a higher IDF. For more details on this methodology, refer to our validation exercise with the World Bank. This distance is determined by the number of overlapping skills (from the skills genome) and the relative importance of each skill to each entity. Entities with a high degree of similarity in their skill sets will have a higher similarity score. The score ranges from 0 (completely different) to 1 (identical). For more on this methodology, see the validation exercise with the World Bank.

- *Popularity:* This indicator indicates how likely it is to move from one occupation to another. For a transition from occupation group A to occupation group B, it is calculated as the percentage of transitions from occupation group A to occupation group B, divided by the percentage of transitions from group A to the most popular alternative occupation, C. The result is then multiplied by 10 to scale the popularity score within a range of 0 to 10. A score of 10 indicates that occupation group B is the most common occupation to transition to from occupation group A.

### 1.2.2. Added value

LinkedIn provided aggregated data on 2024 career transitions, using the platforms own occupation classification (along with a crosswalk to ISCO-08 codes). This data includes the skill similarity metric and the popularity of each transition by country. The countries included are: Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. These are self-reported skills that LinkedIn users have listed in their profiles. This dataset offers a direct and up-to-date reflection of the skills workers perceive as important for their professional identity. Unlike expert-driven skill taxonomies like O*NET or ESCO, which rely on expert assessments to categorize and define the relevance of skills, LinkedIn data captures the skills that workers actively endorse and list on their profiles. These self-reported skills can offer a more dynamic and granular perspective of the labour market, reflecting not only established competencies but also emerging trends and skills that professionals consider important. Furthermore, LinkedIn's skill similarity metric provides valuable insights into the degree to which workers signal their productive capabilities (Spence, 1973). By assessing how closely the skill sets of different occupations or industries align, the metric helps measuring the degree of skill overlap between entities, offering a clearer view of career transitions and skill transferability. This self-reported, real-world data helps to identify how professionals define their skill sets and how these skills relate to other industries or occupations. This contrasts with the more traditional, expert-driven frameworks, which may not capture the same level of real-time, user-generated insight.

In addition to the above, LinkedIn's skill similarity scores can be used as a validation tool to assess the usefulness of O*NET and ESCO in capturing real-world skill alignment and career transitions. By comparing LinkedIn's skill similarity scores with those derived from expert-driven frameworks like O*NET and ESCO, we can validate whether these taxonomies accurately reflect the skills that professionals themselves are reporting and using in their careers. This approach allows for a comparison between theoretical, expert-driven classifications and the dynamic, user-generated skill data from LinkedIn. If LinkedIn's skill similarity scores align with O*NET and ESCO in identifying key skill sets for various occupations or industries, it can confirm the relevance of these frameworks. Conversely, discrepancies between the two could highlight gaps or evolving trends not fully captured by O*NET or ESCO, offering opportunities for refining these taxonomies to better reflect real-world labour market dynamics.

### 1.2.3. Limitations

The main limitation of using LinkedIn data is its aggregation, which can obscure individual-level variations and result in a loss of granular insights. This aggregation may not capture the full spectrum of skills and career transitions at the individual level, potentially overlooking nuances in the reported skill sets of workers. While the skill similarity metric provides valuable insights into the alignment of most relevant skill sets between occupations, industries, or countries, the use of the TF-IDF algorithm relying on the assumption that the most common skills across LinkedIn profiles are less significant. This could result in highly common skills such as Microsoft Word being down-weighted, potentially underestimating their importance for certain occupations or contexts—every rose has its thorn when it comes to this matter. The skill similarity score is thus calculated based on the distance between the skills sets of two occupations. While this offers a broad overview of how closely two occupations align in terms of skill sets, the aggregated nature of the data might obscure more specific, yet critical, skills that distinguish roles within a certain occupation or industry or company size. Consequently, the data may not fully reflect the diversity of skills that individuals possess within the same occupation group, leading to a less complete picture of skill alignment.

Furthermore, the popularity metric, which quantifies the likelihood of transitioning from one occupation to another, is also based on aggregated data. This metric relies on the percentage of transitions between occupation groups, but the aggregation may obscure shifts that are significant for smaller groups or emerging occupations, or for transitions specific to certain company sizes. While the metric offers a useful general overview of transition patterns, it may not fully capture the complexity of career mobility or the dynamics of less common career paths.

Lastly, the calculation of the skills genome is an ongoing and dynamic process. While this aspect is not explicitly detailed in the World Bank reports, it was communicated to us directly. In other words, the list of skills associated with specific occupations can change over time. LinkedIn's skills genome is an experimental measure that is continuously evolving and being updated.

## 1.3. Social media campaigns to collect resumes

To collect resumes, we initiated multiple campaigns to through LinkedIn, X, Google and Meta platforms from Mid-November 2024 to end of January 2025, as well as social media posts on the SkiLMeeT accounts that where shared by the research team. All platforms offer various targeting options, including geographic location, which aligns with our study's focus. Additionally, LinkedIn allows us to target specific job titles relevant to our research objectives . X, google and Meta allow us specify interests of the targeted audience e.g. green and digital solutions, and we used pictures and videos optimized for each platform.

With an investment of EUR 10.000 in the campaign of the first year, we estimated to receive 8.000-10.000 resumes. However, via the campaigns we collected only 22 unique and valid CVs over the course of 2.5 months. As an effort to increase response, multiple measures were implemented:

- Numerous rounds of critical review of participants information for ease of understanding and time efficiency
- Displaying the advertisements and survey in multiple languages (English, Dutch, German, French)
- Increasing trustworthiness of the campaign by building a landing page for participants on the SkiLMeeT homepage (SkiLMeeT)
- Running campaigns from official TNO accounts (branding instead of the newer SkiLMeeT (accounts) and working together with the Marketing & Communication of TNO
- Working together with a Google Ads advisor to optimize the targeting of the campaign

The decision was made to close all the survalyzer links and WeTransfer links as of 31-01-2025.

## 2. References

Spence M (1973) Job Market Signaling. *The Quarterly Journal of Economics* 87(3): 355.

World Bank Document

# List of Tables

Table 1.	Brief overview of the data........ 5

Table 2.	Description of variables in user profile data set........ 6

SkiLMeeT

Follow us:

in linkedin.com/company/SkiLMeeT

bsky.app/profile/skilmeet.bsky.social

twitter.com/SkiLMeeT

www.skilmeet.eu

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under grant agreement No 101132581.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

## SkiLMeeT Partners:

TNO innovation for life

rwi

LISER Luxembourg Institute of Socio-Economic Research

Utrecht University

ibs institute for structural research

Fondazione Eni Enrico Mattei

Institute of Philosophy and Sociology

Bulgarian Academy of Sciences

Aldgate Strategy Group