The crosswalk between ESCO and O*NET

Technical Report – September 2022

Employment, Social Affairs and Inclusion

Acknowledgments

This work is the result of a joint effort between the European Commission, Directorate General for Employment, Social Affairs and Inclusion, and the US Department of Labor/Employment and Training Administration. The work has been led by the ESCO Secretariat, on behalf of the European Commission, and Pamela Frugoli on behalf of the US Department of Labor.

Table of Contents

Table of Figures and Tables

Introduction

The present technical report provides an in-depth explanation of the work performed to produce a crosswalk between ESCO, the European multilingual classification of Skills, Qualifications, and Occupations, developed by DG Employment, Social Affairs and Inclusion of the European Commission, and O*NET, the Occupational Information Network, developed by the U.S. Department of Labor/Employment and Training Administration (US DOL).

In the past, crosswalks were created only sporadically[1], as they require a remarkable effort in terms of time and resources. Today, new technologies such as Machine Learning and Natural Language Processing significantly reduce the effort required to carry out such activities. This report shows the result of the work done to connect the ESCO and O*NET classifications using machine learning models and human validation.

After describing the two classifications, the report clarifies the purpose of the crosswalk, details the methodology used to map the ESCO classification to O*NET, and concludes with the results of the mapping.

The two taxonomies

This section briefly presents the two taxonomies.

The ESCO Classification

ESCO (European Skills, Competences, Qualifications and Occupations) is the European multilingual classification of Skills, Competences and Occupations. ESCO is a European Commission project, run by Directorate General Employment, Social Affairs and Inclusion (DG EMPL).

ESCO works as a **dictionary,** describing, identifying and classifying professional occupations and skills relevant for the EU labour market and education and training. Concepts and the relationships between them can be understood by electronic systems, which allows different online platforms to use ESCO for services like matching jobseekers to jobs on the basis of their skills, suggesting trainings to people who want to reskill or upskill etc.

ESCO provides descriptions of **3008 occupations**, translated into 28 languages (all official EU languages plus Icelandic, Norwegian, Ukrainian, and Arabic).

The aim of ESCO is to **support job mobility across Europe and therefore a more integrated and efficient labour market**, by offering a “common language” on occupations and skills that can be used by different stakeholders on employment and education and training topics.

The O*NET Classification

The O*NET Program is the primary source of descriptive occupational information across the U.S. economy. The Occupational Information Network (O*NET) is developed under the sponsorship of the **U.S. Department of Labor/Employment and Training Administration** (USDOL/ETA) through a grant to the North Carolina Department of Commerce. For more

[1] Crosswalk between the International Standard Classification of Occupations (ISCO-08) and the 2010 Standard Occupational Classification (SOC), approved in July 2012. Technical note available at: < https://www.bls.gov/soc/isco_soc_crosswalk_process.pdf>

information on the O*NET Program, see "About O*NET" at https://www.onetcenter.org/overview.html

Central to the project is the O*NET database, containing hundreds of standardized and occupation-specific descriptors on almost **1,000 occupations** covering the entire U.S. economy. The database, which is available to the public at no cost in English and Spanish, is continually updated from input by a broad range of workers in each occupation.

Every occupation requires a different mix of knowledge, skills, and abilities, and is performed using a variety of activities and tasks. Based on the U.S. Standard Occupational Classification, the O*NET-SOC taxonomy currently includes 923 occupations which currently have, or are scheduled to have, data collected from job incumbents or occupation experts. This linkage between O*NET and the SOC links O*NET descriptive information to quantitative information on occupational employment, wages, and other variables from surveys conducted by the U.S. Department of Labor, Bureau of Labor Statistics. For more information on the O*NET occupational classification see: "The O*NET-SOC Taxonomy" https://www.onetcenter.org/taxonomy.html#latest and also the report: Updating the O*NETSOC Taxonomy: Incorporating the 2018 SOC Structure. The 2018 Standard Occupational Classification (SOC) system is a United States federal statistical standard used by federal agencies to classify workers into occupational categories for the purpose of collecting, calculating, or disseminating data. All workers are classified into detailed occupations based on the occupational definition[2].

Purpose of the crosswalk

Creating a crosswalk between the two taxonomies is of great importance to support interoperability between two labour market standards used by a multitude of public and private stakeholders to provide services such as job matching, upskilling and reskilling, matching people with the right training opportunity, statistical analysis. A bridge between ESCO and O*NET is particularly useful for researchers and public officials, who are involved in developing policies to improve the labour market, or carrying out research on topics related to workforce and education.

An official mapping, co-created and approved by the owners of both classifications, guarantees a high level of quality and reliability and therefore would encourage its use by a wide range of organisations, including those that do not have the resources to make such a crosswalk or ensure the reliability and quality of the data.

Methodology

The mapping of the two classifications was performed by combining the use of artificial intelligence (AI) techniques with human validation. Such synergy assured higher accuracy and consistency of the results, while reducing the manual effort in terms of working hours.

[2] The first six-digits of every eight-digit O*NET code indicates the corresponding SOC code. If the 7th and 8th digits of the O*NET code after the decimal, are .00 then the O*NET and SOC occupation are identical. If the numbers after the decimal are greater than zero that indicates an additional detailed breakout specific to the O*NET taxonomy.

The methodology adopted is represented in Figure 1 and described below.

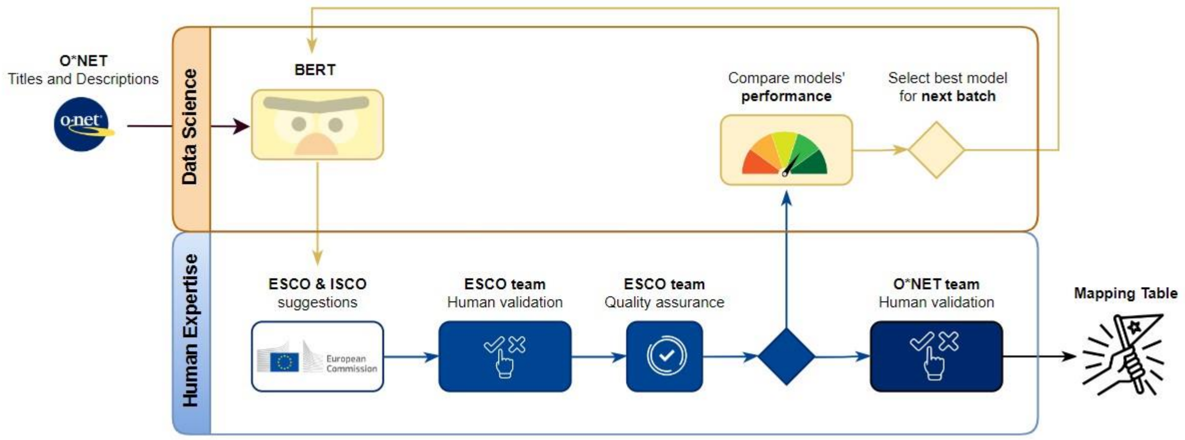

*Figure 1: ESCO-O*NET crosswalk methodology*

O*NET occupations were divided into three groups through stratified random sampling based on SOC groups. Each group was treated as a single batch of work. Batches were processed in different periods starting in mid-2021 and finishing in mid-2022. Details are provided in Table 1.

The European Commission developed AI models that match O*NET occupations (input) to ESCO occupations (output) based on semantic textual similarity. The models were trained using labour market expert feedback, national taxonomies, Qualification Data Register[3] qualifications, and online job vacancies. Based on the best-performing machine learning model, ten ESCO occupations were suggested for every O*NET occupation. These ten suggestions represented the ESCO occupations with the highest-ranking scores and were considered as semantically most similar to the respective O*NET occupation by the model. To further ease the validation process, the team included the results of a past mapping between the International Standard Classification of Occupations (ISCO-08) and SOC carried out in 2010 by the US DOL.

Next, the ESCO team proceeded with human validation: the occupations were equally distributed among the team members, who proceeded to determine the relationship between each suggested O*NET-ESCO occupation pair. This was done following a pre-established set of rules which defined a range of possible relations between concepts and the detailed information about those concepts (i.e. descriptions, tasks, job titles, etc.). The section *Mapping guidelines* of this report provides more details regarding the set of rules. Validations went through a quality assurance process, where a different validator checked the established relations and, in case of disagreement, a third validator was involved. This is further explained in the *Quality assurance section.*

As a result of the validation process, ESCO occupation suggestions from the machine learning model were split in two sets of results. One set consisted of suggestions considered to be *relevant* (exact, narrower, broader, and close match) as obtained through human validation. These results were shared with the O*NET colleagues for a final step of manual validation. The second set of results included both the *relevant* and the *less relevant*

[3] https://europa.eu/europass/qdr/

suggestions (related and wrong matches), which were used to compare different machine learning models and select the best performing one for processing the next batch of O*NET occupations.

	O*NET	ESCO	ISCO
Classification version (*year*)	O*NET-SOC 2019 (*2019*)	ESCO v1.1.0 (*2022*)	ISCO-08 (*2007*)
Occupations mapped *first batch*	199	1,231	143
Occupations mapped *second batch*	351	1,683	196
Occupations mapped *third batch*	373	1,098	206
Occupations mapped *orphans batch*	35	652	0
Unique occupations mapped	958	2997	352
Occupations excluded from the crosswalk	58	11	267
Total unique occupations	1,016	3,008	619
Number validators	1	6	6 (ESCO validators)

Table 1: Details on the work distribution

Note: Validation batches originate from sampling O*NET occupations in three distinct groups. The same ESCO occupation (output from the model) can be mapped in different batches for a different match type. The same O*NET occupation (input for the model) cannot be mapped in more than one batch.
Note: While the O*NET classification is composed of 1,016 occupations, 58 occupations were excluded from this exercise: 16 armed forces occupations (SOC group 55) were removed because of incomparability between the European and U.S. market, and 42 O*NET occupations characterised by the *All other* label. *All* other occupations were taken into account only in those cases where the ESCO occupation remained unmatched (i.e. when running the *orphans batch*).
Note: 11 ESCO occupations were excluded from the crosswalk because related to the armed forces roles.
Note: The baseline version of O*NET selected is the one reported in the table, but smaller changes may apply in the final crosswalk table as quality assurance checks considered the online version of O*NET, which is updated to the latest changes.
The table was updated in December 2022 to reflect corrections.

The division of the workload into batches allowed to apply different machine learning models in the three phases of the process. As such, working with batches represented an efficient approach to more quickly improve the quality of the suggested ESCO occupations over time. The improvements resulted from evaluating various models, that used different training data types and quantity, finetuning strategy, and training time. The next section provides more details.

Artificial Intelligence model

The European Commission is using data science technologies to support the work done by experts to maintain and improve ESCO, and to ease the use of the classification for implementers. This includes developing strategies to connect external information concerning the labour market to ESCO concepts. The use of machine learning models to connect external occupations, job titles and supporting information to an ESCO occupation was already illustrated in different publications on the ESCO portal[4].

To establish a crosswalk between the European and U.S. classification of occupations, the ESCO Data Science Team focused on developing an approach to detect the most semantically similar ESCO occupations for an O*NET occupation. Semantic similarity between occupation concepts is defined over the closeness at meaning or semantic level between O*NET job titles and descriptions, and ESCO preferred terms, alternative terms and descriptions. Semantic similarity plays a significant role in various natural language processing (NLP) tasks as it allows to assign a score, representing likeness of meaning, to

[4] *The role of contextual information when connecting data to the ESCO Occupations Pillar using Artificial Intelligence*, ESCO. Available at: https://esco.ec.europa.eu/en/about-esco/data-science-and-esco/rolecontextual-information-when-connecting-data-esco-occupations-pillar-using-artificial (Accessed: 2022)

the relationship between two textual items. For the purpose of generating ESCO occupation suggestions, the approach was to apply transformer language models to encode definitions of occupations in embeddings and then to the use the cosine similarity metric to compute a similarity score.

This project demonstrated the efficacy of the Bidirectional Encoder Representations from Transformers (BERT) language model to process labour market information. BERT is a language representation model developed by the Google AI Language group[5]. The ESCO Data Science Team fine-tuned BERT through multi-task learning while adding an additional linear layer to the original BERT model. Different strategies to optimise model performance (i.e. a model that, for one O*NET occupation, assigns the higher score to the most similar ESCO occupation) included using different optimisation methods, loss functions, combinations of the input information and training iterations.

After validating each batch of O*NET occupations, outputs from different models were compared with the expert validations, to detect to what extent the score assigned by the model to the relationships would reflect the types of relationship assigned by the validator. The model performing the most similar as compared to the human validation was then selected to predict the suggestions for the following batch. Tables 2, 3, and 4 list the details of the selected model.

	Model 1	Model 2	Model 3
Model	Fine-tuned bert-base-uncased	Fine-tuned bert-base-uncased	Fine-tuned bert-base-uncased
Source embedding	Concatenate O*NET title and description	Separate embeddings for O*NET title and O*NET title + description	Separate embeddings for O*NET title and O*NET title + description
Target embedding	Average of concatenated permutations of ESCO (non-)preferred term(s)	Separate embeddings for ESCO (non-)preferred term(s) and description	Separate embeddings for ESCO (non-)preferred term(s) and description
Scoring	Cosine similarity (cs)	0.4 * max(cs(ESCO pt/npts, O*NET title)) + 0.3 * median(cs(ESCO pt/npts, O*NET title)) + 0.3 * (cs(ESCO desc, O*NET title+desc))	0.3 * max(cs(ESCO pt/npts, O*NET title)) + 0.15 * median(cs(ESCO pt/npts, O*NET title)) + 0.55 * (cs(ESCO desc, O*NET title+desc))
Batch size	24	24	24
Model updates	2,395,282	2,395,282	2,826,735
Training size	25,376,400	25,376,400	29,756,109

Table 2: Details on the AI model employed for the occupation mapping

Note: the ESCO-O*NET crosswalk compares two structured datasets with a controlled lexicon. Results may differ when processing different data types, such as unstructured descriptions extracted from job vacancies.

Top k accuracy metrics for exact match relation at different stages of the process. For example, model 1 was developed and suggestions were generated for batch 1. After manual validation of those batch 1 suggestions, 70.43% of the O*NET occupations having an exact match (i.e. 115 out of 199 occupations) had the exact match suggested at position 1 (i.e. 81

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2016). *Bert: Bidirectional Encoder Representations from Transformers.*

out of 115 occupations). For the same batch all exact matches appeared in the top 10 of suggestions (i.e. 115 out of 115 occupations).

Batch	Model 1	Model 2	Model 3
O*NET Occupations	115	199	177
Top 1 Accuracy	0.7043	0.8291	0.8475
Top 2 Accuracy	0.8261	0.9196	0.9153
Top 3 Accuracy	0.9043	0.9497	0.9605
Top 5 Accuracy	0.9478	0.9799	0.9774
Top 10 Accuracy	1.0	1.0	1.0

Table 3: Details on the performance of the AI models employed for the occupation mapping (exact matches)

Top k accuracy metrics for the highest ranked exact, broad, narrow or close match relation at different stages of the process. For example, model 3 was developed and suggestions were generated for batch 3. After manual validation of those batch 3 suggestions, 90.96% of the O*NET occupations having either an exact, broad, narrow or close match (i.e. 365 out of 373 occupations) had either of those matches suggested at position 1 (i.e. 332 out of 365 occupations). For the same batch 99.45% of the highest ranked exact, broad, narrow or close match appeared in the top 5 of suggestions (i.e. 363 out of 365 occupations).

Batch	Model 1	Model 2	Model 3
O*NET Occupations	196	344	365
Top 1 Accuracy	0.8469	0.8895	0.9096
Top 2 Accuracy	0.9439	0.9419	0.9671
Top 3 Accuracy	0.9643	0.9593	0.9836
Top 5 Accuracy	0.9745	0.9767	0.9945
Top 10 Accuracy	0.9949	1.0	1.0

Table 4: Details on the performance of the AI models employed for the occupation mapping (exact, broad, narrow, close match)

Mapping guidelines

The types of mapping relations identified correspond to those used during the mapping of national classifications to ESCO[6]. These definitions have been adapted to the scope of this exercise. The relations and their definitions are listed below (see Figure 2 for a visual explanation):

- **Exact match:** A concept in O*NET covers the same scope as a concept in ESCO and vice versa. The scope is defined by 1) the jobs that fall under / are associated with this occupation and 2) the tasks, skills, activities, and knowledge linked with an occupation. Each O*NET occupation can have only one exact match. Example: O*NET Mechanical Engineers is an exact match of ESCO Mechanical Engineer

- **Narrow match:** A concept in O*NET is more specific than a concept in ESCO, as it covers only a fraction of its scope. This O*NET occupation can be considered as a narrower occupation of the ESCO occupation proposed by the model. Example: O*NET Environmental Restoration Planners is a narrow match of ESCO Environmental Programme Coordinator
- **Broad match:** A concept in O*NET is more general than a concept in ESCO, as it covers its full scope and more (i.e., additional occupations). This O*NET occupation can be considered as a broader occupation of the ESCO occupation proposed by the model. Example: O*NET Art, Drama, and Music Teachers, Postsecondary is a broad match of ESCO Drama Teacher Secondary School.
- **Close Match:** The occupation proposed by the model is not an exact, broad, or narrow match, but it covers similar or a large part of the scope of the O*NET occupation (in other words, there is overlap between the two occupations. This case may occur when a classification is based on a dimension or perspective that is not reflected in the other classification (i.e., one is based on a product and the other on a manufacturing process/technique). In other words, a concept can be identified that is very similar, but not identical. Example: O*NET Surveying and Mapping Technicians is a close match of ESCO Geographic Information Systems Specialist.

In order to further enhance the performance of the AI model and the precision of the mapping, two additional relations have been defined, which are not included in the official mapping but have been fundamental to clarify the scope of the Close Match relation and to analyse the outputs of the machine. The supplementary relations are:

- **Related Match:** The occupation proposed by the model is different (has a different scope) from the O*NET occupation, but such a suggestion is still reasonable, there is a degree of relatedness. For example, the two occupations may be characterized by a similar career path, even if moving from one occupation to another would require limited additional training (for example, this excludes career pathways between occupations such as nurses and doctors). Example: O*NET Nuclear Technicians is a related match of ESCO Radiation Protection Officer.
- **Wrong match:** The relation does not correspond to any of the above and therefore the concepts are not related, and there should not be this suggestion. Example: O*NET Billing and Posting Clerks is a wrong match of ESCO Meter Reader.

[6] The Regulation (EU) 2016/589 provides for the obligation for the members of the EURES network to map their national classifications to ESCO or to adopt ESCO, in order to enable the exchange of job vacancies and CVs across borders through the EURES platform. The result of this exercise is available at this page.

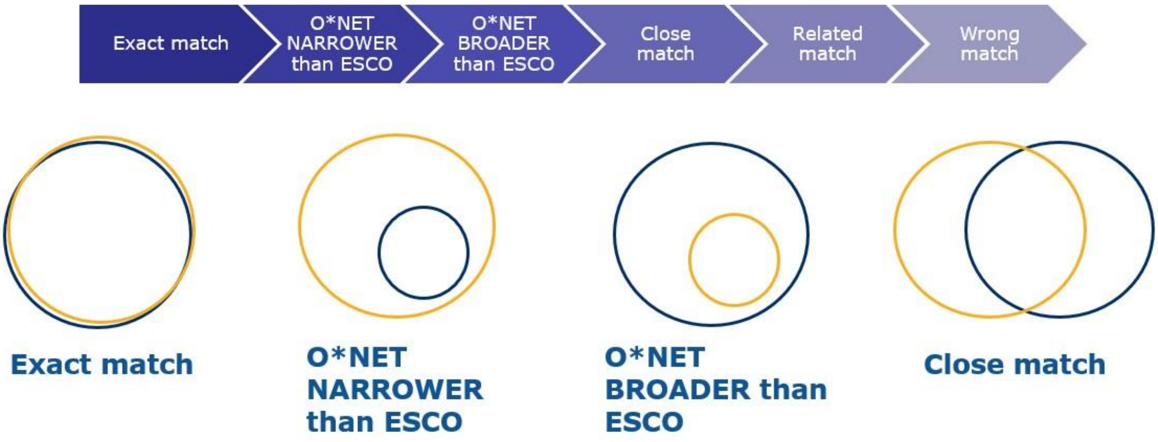

Figure 2: Types of mapping relations

The steps followed to validate the relations between concepts are the following:

1. Read the O*NET title and description,
2. Proceeding row by row, read the corresponding ESCO preferred term and description,
3. When necessary, consider all the data that can be gathered from both O*NET and ESCO, such as alternative labels and skills for ESCO, and tasks, skills, and sample of reported job titles. Moreover, look for similar occupations (e.g. if the O*NET title refers to a managerial occupation, check if for the same area of work O*NET includes assistant or operative occupations),
4. Where available, compare the ISCO group of the ESCO occupations suggested with the ISCO group resulting from the ISCO-SOC crosswalk, and
5. Establish the type of relation between the two terms and mark it in the corresponding column.

Due to the differences between the North American and European labour markets, it was not possible to identify in the ESCO classification exact matches for all the selected O*NET occupations. However, as ISCO represents the hierarchical structure of ESCO, the O*NET concepts have also been mapped to ISCO level 4.

Such mapping yielded two types of relations between O*NET and ISCO:

- **Exact Match:** defined as the exact match between O*NET and ISCO;
- **Best Match:** the closest ISCO concept to the O*NET occupation, includes cases when O*NET is broader, narrower or a close match to ISCO.

The mapping to ISCO followed the same process as the mapping to ESCO. The machine learning model predicted the 5 closest ISCO concepts for each O*NET occupation, and relations were established after human validation, also considering the SOC-ISCO mapping of 2010 and the ISCO codes of the ESCO occupations found narrower, broader, or close to the O*NET concept during the first phase of the mapping.

Validation of the orphans

Once the mapping from O*NET to ESCO was complete, it was detected that 652 ESCO occupations were missing from the crosswalk. For such concepts, no type of relation was established - this can be attributed in large part to the higher level of granularity of some ESCO occupations. In order to ensure the completeness of the mapping, the ESCO team manually analysed the “orphan” occupations to link them to the appropriate O*NET correspondent.

The ESCO concepts mapped as *related* to at least one O*NET occupation were not considered orphans. As a consequence, such concepts are not included in the official mapping relations, but they are available in the additional mapping table published in the ESCO portal, where *related* matches are listed. The review of these occupations is foreseen for the update of the mapping.

Quality assurance and process

The quality of the results is guaranteed by the process of quality assurance (QA): after the first training meeting, each team member mapped an initial batch of 5 O*NET occupations. These mappings have been discussed during the second training meeting, where methodology and guidelines have been refined. The team proceeded to validate the entire file, and as doubtful cases emerged, they were discussed bilaterally or with the whole team. Every time the project team completed the mapping of a batch, a smaller group performed in-depth checks of the validation results and verified the cases of disagreement with the people responsible for the validation. Moreover, the QA team verified that the exact matches were only one-to-one, and that in case of multiple narrow matches the mapping still made sense. Few cases were accepted even if not following the validation rules, as they were considered justifiable exceptions.

To ensure alignment between the ESCO and O*NET perspectives, after the first batch was mapped, the two teams had a workshop where they discussed open questions and doubts related to the mapping. Finally, after every batch the O*NET team performed a comprehensive review and suggested potential revisions to provide additional quality assurance; this provided fundamental insights into the O*NET classification.

Results

The work presented in this report led to the publication of a crosswalk between the ESCO and O*NET classifications for occupations. All the ESCO occupations are mapped to at least one O*NET occupation, based on different types of relations, as defined in the *Mapping guidelines.* The range of relations eases the process of comparisons of the two classifications, which differ by number of concepts and granularity.

Looking in detail at the result, validators agreed on 542 exact matches between the two classifications, of which 499 are between ESCO occupations and O*NET occupations, and 43 are between ISCO groups and O*NET occupations. The number of narrow matches is reasonably lower, as it appears that about 227 times an O*NET occupation has been considered more detailed than an ESCO occupation. On a similar line, a higher number of broad matches is found, connecting 2,066 times an O*NET occupation with a more detailed ESCO occupation. These differences are justified by the difference in size of the two classifications, where the European classification is three times as big as the U.S. counterpart. In many cases, occupations between the two classifications cover a similar scope, but cannot be considered as identical, more general, or specific. This is the case for the 1,442 close relations flagged.

The total number of validations amounts to 5,175 matches, or 7,385 if considering *related* matches.

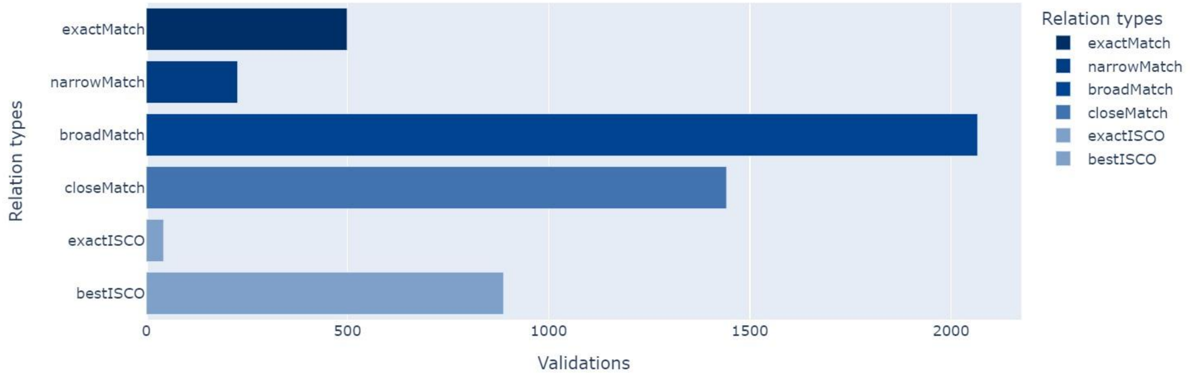

Figure 3: Distribution of validations by relation type

Figure 4 illustrates the distribution of the different relation types among the first level of the ISCO occupation hierarchy, which is the starting point of the structure of the ESCO occupation pillar. The distribution of relation types is represented in percentage points, to help understanding to what extent ESCO concepts mapped under the same occupation group coincide with O*NET occupations. The two classifications seem to be more harmonised around clerical occupations and elementary occupations, where examples are the ESCO payroll clerk and the O*NET Payroll and Timekeeping Clerks, the ESCO domestic cleaner and the O*NET Maids and Housekeeping Cleaners. Occupations concerning skilled agricultural, forestry and fishery workers, as well as service and sales workers, and managers seem to be more detailed in the ESCO classifications compared to the O*NET classification. Differently, the O*NET classification is more detailed than the average compared to ESCO around the field of professional occupations.

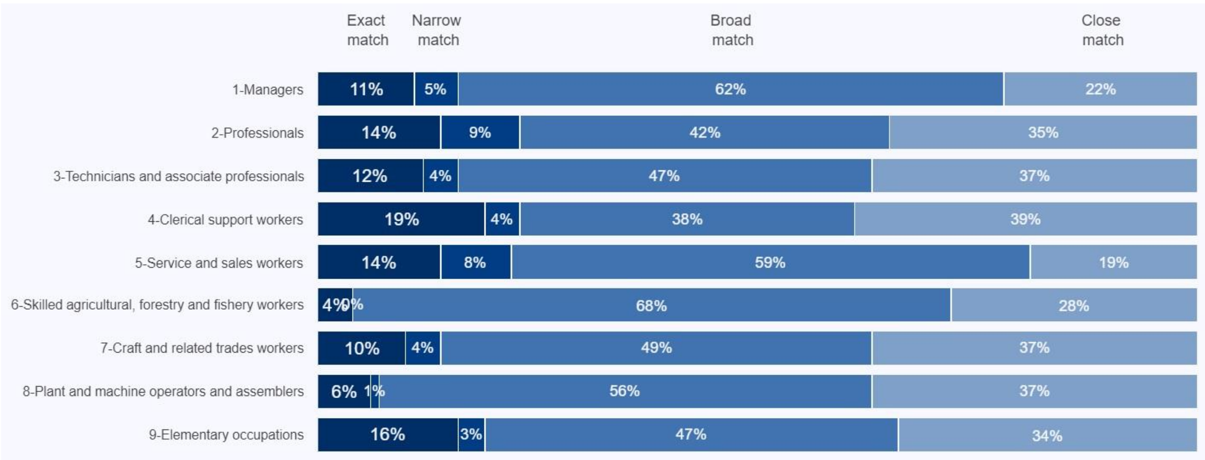

Figure 4: Distribution of validations by ISCO group

When comparing the two classifications, especially if using percentage values, it is important to remember that each occupation group has a different number of occupations, where differences in percentages may not well depict variations in the raw number of occupations.

Conclusion and next steps

The crosswalk between ESCO and O*NET serves the purposes of job-matching platforms, researchers, public officials, and other stakeholders to better understand the labour market, provide skills-based services, or formulate data-driven policies.

The labour market is in continuous evolution, and the ESCO and O*NET taxonomies are updated regularly. This implies that the work on the crosswalk will continue: the European Commission aims at revising it every time a new major version of ESCO will be released.

Finally, it must be noted that methodology and models developed for the crosswalk between ESCO and O*NET have the potential to be employed to map ESCO with many other non-EU classifications, which is extremely useful to support interoperability between different labour market standards.

Supporting Material

The crosswalk between ESCO and O*NET can be accessed from both the ESCO and O*NET portals.

In the ESCO portal it is available in the section Other Crosswalks. Additionally, the Data Science Blog provides details regarding the methodology.

In the O*NET Resource Center portal the crosswalk is available in the Crosswalk Files section: https://www.onetcenter.org/crosswalks.html. ESCO will also be available as a crosswalk search option in the Crosswalks section of the O*NET Online homepage and will be available to developers via O*NET OnLine web services.