EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR EMPLOYMENT, SOCIAL AFFAIRS AND INCLUSION

Labour Mobility and International Affairs
Labour Mobility, Public Employment Services, ELA

Brussels
EMPL.E.1

ESCO MEMBER STATES WORKING GROUP
FOCUS MEETING
14 SEPTEMBER 2022

Subject: MWSG focus meeting concept note: Additional clustering of ESCO concepts

CONTENTS

1) PURPOSE OF THIS MEETING

The Commission invites the input from members and observers of the ESCO Member States Working Group on the opportunity of additional collections of skills and/or occupation concepts in ESCO, with a particular focus on the technical implementation.

ESCO v1.1.0[1] consists of two pillars, the occupations pillar and the skills pillar. The two pillars are organized based on a top-down approach, where concepts are mapped to pre-established categories, and follow a mono-hierarchical architecture. The current hierarchical structures are the result of testing and rich consultation with taxonomy experts. While they have proven successful based on the feedback of experts, implementers and researchers, the Commission registered increased interest in additional clusters of ESCO concepts, in particular for skills and competences.

Rather than replacing them, alternative structures would complement the current ESCO hierarchies. Building on the ambition to continuously evolve ESCO into a connected knowledge graph, the Commission is now exploring the possible ways forward.

The paragraphs below describe in detail the alternatives scoped out as part of this exercise. As for the skills classification, **the main scenario in focus is a clustering-driven hierarchy** that would follow a bottom-up approach. The **main alternative to the ESCO occupations hierarchy is a NACE-based occupations classification** (top-down approach). However, the Commission seeks input from Member States, experts and other relevant stakeholders based on the subject matter more broadly, going beyond these two scenarios.

2) BACKGROUND: CURRENT ESCO HIERARCHY STRUCTURE

The ESCO database in its current form consists of two pillars: the skills and the occupations pillars. The architecture underlying each of these pillars follows the principle of mono-hierarchy where each concept appears in the hierarchy only once.

The occupations pillar

Occupations in ESCO are structured through their mapping to the International Standard Classification of Occupations (ISCO-08). ISCO-08 provides the top four levels while ESCO occupations provide the fifth and lower levels. Each ESCO occupation is assigned to only one ISCO-08 unit group (even if they are not directly related to it, e.g. if they are at level six or seven).

Figure 1 below illustrates the structure of the occupations pillar.

[1] A new minor version of ESCO, ESCO v1.1.1, will be released during the course of September.

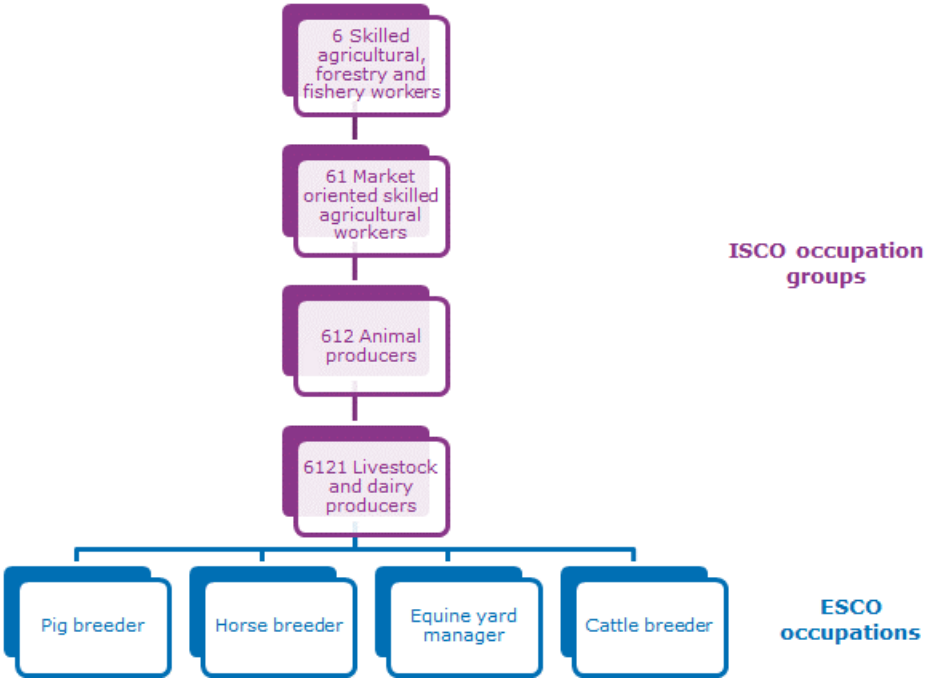

Figure 1 Structure of the occupations pillar

The skills/competences pillar

The 13890 knowledge and skills concepts are classified under the ESCO skills hierarchy, a single all-embracing hierarchical framework composed of four different sub-classifications:

- the skills/competence hierarchy (S)
- the knowledge hierarchy (K)
- the language hierarchy (L)
- the transversal skills and competences hierarchy (T).

The Table 1 below presents the underlying structure per each of the four sub-hierarchies.

Concepts	Sub-hierarchies
Knowledge	International Standard Classification of Education, Fields of Education (ISCED-F).
Languages	A three-levels structure building on the pre-existing language hierarchy from ESCO v1.0, distinguishing between languages and classical languages.
Skills	A structure of a 3-level skill hierarchy, designed by taxonomy experts and building on the Canadian Skills and Knowledge glossary and O*NET IWAs.
Transversal skills	An expert-created 2-level structure with six headings: core

	skills and competences, thinking skills and competences, self-management skills and competences, social and communication skills and competences, physical and manual skills and competences, life skills and competences.

Table 1 Building structures of the ESCO skill pillar

The primary consideration in the definition of the groups was the mobility between occupations (skill transferability). For this reason, the groups of the skills and competences hierarchy were designed to be as homogeneous as possible in relation to at least one of the following characteristics:

a. Tools and equipment used.

b. The type of object on which the work is performed.

c. The function or outcome of the task or activity.

The choice of the characteristic that should take priority over the others depends on its impact on occupational mobility. For example, there are cases where the type of object on which the work is performed impacts more decisively on the degree of transferability, e.g. live animals or plants or human beings. In other cases, being able to use a particular type of tool may assist more in occupational mobility; therefore, it is this characteristic that takes priority.

Figure 2 below illustrates the structure of the skills/competences sub-hierarchy.

Figure 2 Structure of the skills/competences sub-hierarchy

3) OPTIONS FOR AN ALTERNATIVE STRUCTURING OF THE ESCO SKILLS PILLAR

At present, the ESCO skills classification is structured as a taxonomy, like most of the existing occupations and skills classifications. As such, it is based on a hierarchical structure composed of parent-child relations, which connect the individual concepts across different levels.

There are two main approaches to building a classification: top-down and bottom-up strategies. Under the top-down approach, concepts are mapped to pre-established categories, whereas under the bottom-up, the categories of the hierarchy are formed based on the nature of the concepts that will be mapped.

Top-down approach: clustering of ESCO skills based on existing frameworks

In the top-down approach, existing and recognised frameworks are leveraged to defining structures. This is partially the case of the current ESCO hierarchy, which uses the ISCED-F classification to group knowledge concepts.

In the past, the Commission tested through a series of pilots the usability of existing frameworks for structuring the skills pillar of ESCO. The structural approaches explored were the use of DISCO (European Dictionary of Skills and Competences), Kompetenzklassifikation (the German skills classification), NACE (Statistical Classification of Economic Activities in the European Union), ISCED-F (International Standard Classification of Education - Fields of Education and Training), and CPA (Statistical classification of products by activity). As neither of these classifications was suitable as a single organizing principle for a mono-hierarchy, the Commission invited a group of ontology experts to design an embracing framework to group the over 13000 skills and knowledge concepts part of ESCO V1.0. The ESCO skills hierarchy was finally implemented with the release of ESCO v1.0.5.

In the light of this finding, the bottom-up approach (discussed in more detail below) is prioritized for the current scoping out of opportunities for the hierarchical evolution of ESCO.

Bottom-up approach in ESCO: data-driven classification

Under the bottom-approach, it is the data that forms the building foundation of the hierarchy. In other words, the final taxonomy structure is derived from its content, relying on similarity of the concepts in a database. Use of the bottom-up approach has been recently on the rise thanks to the rapid growth of skills intelligence in recent years. Based on ESCO's research, there are three types of strategies to how the similarity can be determined that could be leveraged for alternative ESCO skills hierarchy: using *co-ocurrence of concepts, using semantic similiarity, or a hybdrid model* that combines the two. These strategies are elaborated in more detail below, as well as their advantages and challenges in the context of ESCO.

1. *Skill concept-level co-occurrence driven clustering*

A well-known approach to group together skills is based on their level of co-occurrence in the labour market.

Anderson[2] constructed a graph where skills are nodes and different skills are connected via an edge if a worker has both skills. Each edge in the graph is assigned a weight based on its frequency of co-occurrence. In addition, Anderson combined this network with data on skills requested by employers, as for example from an online job marketplace or job board. Skills are connected if both are required by the employer for a specific job and a weight quantifies level of co-occurrence. Both graphs combined provide a more complete overview to visualise the relationships between skills. Next, an algorithm from the field of network analysis that extracts communities from large networks is used to identify strongly connected groups of skills in the combined network. By identifying sub communities of skills that are often used together, a clustering is obtained.

A prerequisite for this methodology is the availability of a representative amount of skill concept-level co-occurrence data from the labour market as present in online job vacancies or curriculum vitae for example. This might include the need for an approach to link free text data sources to skill concepts.

A significant advantage of the approach is that it is very transparent in terms of similarity metric: skills have a higher likelihood to be part of the same cluster if they are more often used together compared to other skills.

2. *Semantic relatedness-driven skill clustering*

Co-occurrence of skill concepts is a powerful approach, but it does not directly take advantage of the actual content of skills (i.e. the words in the preferred term, alternative terms, description). As an alternative, Gallagher et al.[3] proposed to use a language model to measure *semantic similarity between skill concepts based on their surface forms as they appear in online job advertisements*. This relates to the use of language.

Language models are usually based on what is called the "distributional hypothesis", i.e. linguistic items with similar distributions have similar meanings. Practically, this means that words that share the same context (e.g. neighbouring words), are more likely to be related. This approach has been demonstrated to be very successful for finding semantically related items. However, Jatnika et al.[4] showed that the distributional hypothesis can lead to models capturing semantic relatedness rather than semantic similarity. Reimers and Gurevych[5] further optimised a BERT-based language model by

[2] Katharine A. Anderson, Skill networks and measures of complex human capital, Proceedings of the National Academy of Sciences, 114(48), 12720-12724, 2017

[3] Elizabeth Gallagher, India Kerle, Cath Sleeman, George Richardson, A New Approach to Building a Skills Taxonomy, Economic Statistics Centre of Excellence (ESCoE) Technical Reports ESCOE-TR-16, Economic Statistics Centre of Excellence (ESCoE), 2022.

[4] Derry Jatnika, Moch Arif Bijaksana and Arie Ardiyanti Suryani, Word2Vec Model Analysis for Semantic Similarities in English Words, Procedia Computer Science, 157, 160-167, 2019.

[5] Nils Reimers and Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 11, 2019.

finetuning and evaluating on a semantic textual similarity task. This language model produces vectorial output for text input, where close vectors are more likely to be semantically similar.

Gallagher et al. used the model as developed by Reimers and Gurevych to compute vectors for phrases (representing skills), which were extracted from a large set of online job advertisements. After dimensionality reduction, the skill vectors are clustered in different stages, resulting in a 4-level hierarchy with 6685 skills at the lowest level. The authors describe that this can be used as a basis for further manually developing a skills taxonomy.

The advantage of the approach by Gallagher et al. is that it can directly leverage the state-of-the-art models for semantic relatedness and apply them to the words in skill phrases. The disadvantage is that these language models are not specifically developed on labour market data and do not take advantage of data sets that contain inherent knowledge about skill usage at the concept level (e.g., skill co-occurrence).

3. Hybrid methodology for skill clustering

Recently, multiple studies illustrated the combined use of state-of-the-art language models with skill concept-level co-occurrence data. Decorte et al.[6] and Zbib et al.[7] extracted skills from online job advertisements in order to create skill co-occurrence data. They applied the principle of distributional hypothesis to these co-occurrence datasets in combination with BERT-based language models. Both studies generate vectorial representations for skills derived from skill co-occurrence data. These resulting skill vectors can then be clustered together, similar to the approach in Gallagher et al.

Combining state-of-the-art language models with skill-concept co-occurrence data seems to be a promising approach. For instance, this allows combining skill representations obtained via language models for measuring semantic similarity with skill representations as obtained from co-occurrence data that capture patterns in labour market.

4) OPTIONS FOR AN ALTERNATIVE STRUCTURING OF THE ESCO OCCUPATIONS PILLAR

Similarly to the skills classification, both of the approaches (bottom-up and top-down) could be envisioned for an alternative classification of ESCO occupations. While the information above on the bottom-up approach can be also applicable to the occupations structuring, in this section, the focus is placed onto another possible classification strategy envisioned for ESCO: a NACE-based occupations hierarchy.

[6] Jens-Joris Decorte, Jeroen Van Hautte, Thomas Demeester and Chris Develder, JobBERT: understanding job titles through skills, FEAST, ECML-PKDD 2021 Workshop, Proceedings, 2021.

[7] Rabih Zbib, Lucas Alvarez, Federico Retyk, Rus Poves, Juan Aizpuru, Hermenegildo Fabregat, Vaidotas Simkus and Emilia García-Casademont, Learning job titles similarity from noisy skill labels, Unpublished manuscript, 2022.

A recent survey conducted by the ESCO Secretariat showed how the majority (52%) of ESCO implementers already uses ESCO in combination with NACE. Building on this finding, the Commission plans to publish a mapping between ESCO occupations and NACE classes. This mapping has the potential to lead to an additional clustering of occupations, should Member States and ESCO implementers flag a business need.

As NACE covers the entire economic activity of the European Union and it is used by all major labour market players, there is a strong argument to use the classification as the building principle for an alternative ESCO occupations taxonomy. This would be a top-down approach where concepts are assigned to predefined categories, and it follows a poly-categorical structure, as it can be expected that a large number of ESCO occupations would occur as relevant for multiple fields of economic activity.

To connect ESCO occupations to the NACE classification, the ESCO Secretariat is designing a methodology that combines the use of artificial intelligence algorithms *and* the validation of experts.

More specifically, the ESCO Secretariat is collecting online job vacancies published by European Public Employment Services (PES) in the EURES portal, which are tagged with both ESCO occupations and NACE sectors.

First, running a statistical analysis on this data allows investigating the existing connections made by EURES users of NACE sectors with regards to ESCO occupations. The expected result is a contingency table to highlight NACE sectors mapped to a higher number of occupations, or more frequent occupation-sector combinations, or significant differences among users.

This information is then further cleaned and structured, and it will be used as training data for a machine-learning model, which is expected to learn to predict the most suitable NACE sector, given an ESCO occupation. The model is needed to remove noise, handle missing information and conflicting combinations.

Once several models are drafted, their quality will be tested using a past ESCO-NACE crosswalk which was developed by the ESCO Secretariat in 2011 and further manual quality checks. This would then lead to a comparison of the performance of different models, resulting in the choice of the most accurate model. To conclude, this model would be run with possible improvements and its results will be manually checked.

Apart from the work needed on the model development, other challenges must be considered for this work. Introducing an alternative to ISCO will have technical implications to the current structure of the ESCO data model. Another factor to consider is NACE's limited time-proofness, as it is revised only once every 15 years, which may not provide for a sufficient timeline in the context of the ESCO database to remain up to date with the most recent trends in the labour market.

5) ESCO-SIDE GUIDING THOUGHTS

The following questions are aimed at guiding the discussion between the experts during the thematic focus meeting:

What are the alternatives to the ESCO skills/competences and occupations hierarchies?

What is the value added of introducing alternative hierarchies? What factors should be kept in mind when introducing alternative ESCO hierarchies?

How to strike balance between technical feasibility and value added when selecting alternative classification of ESCO?

What are the benefits and pitfalls of mapping our occupations/skills hierarchies to NACE?

Is data-driven hierarchy feasible for the ESCO taxonomy?

Should data-driven ESCO classification build on vacancy data, semantic similarity of ESCO concept labels and descriptions or another dataset?

What type of ESCO implementers and users would benefit from the different classification alternatives proposed?

Mono vs polycategorical hierarchy? What should the Commission keep in mind when introducing such alternative hierarchies (e.g. issues related to implementation, implications for maintenance)? What are the technical implications of polycategorical hierarchy?

Are there any skill ontologies that are user-friendly and fit for similar purpose as ESCO?