Increasing quality of web-based data: Human Role in production of Consistent Labour Market intelligence

Vladimir Kvetan[1], Jiri Branka[1]

[1] European Centre for the Development of Vocational Training (Cedefop), Thessaloniki, Greece

Abstract

In today's dynamically changing labour markets, real-time skills intelligence supported by official statistics is pivotal for shaping effective employment and education policies. Tapping up on web data creates a powerful avenue to understand employer demands swiftly. A collaborative effort by Cedefop and Eurostat within the Web Intelligence Hub (WIH) focuses on utilising web-derived insights to develop official skills statistics. However, the vast linguistic diversity challenges this data's consistency and cross-country comparability. This highlights the vital need for human input in this process.

This article draws on WIH's work to critically examine the abilities of automatised tools to extract coherent and cross-country comparable information on skills from web-based data. The main departing point of this work is that while the European multilingual classification of Skills, Competencies and Occupations (ESCO) provides a comprehensive skill framework, diverse linguistic representations cause challenges to accurate and comparable skills extraction. The article presents how human intelligence was used to refine and standardise skill extraction.

Building on WIH activities to introduce consistent and cross-language comparable extraction of skills terms, this piece advocates for the pivotal role of human judgment and expertise in enhancing the precision and uniformity of data extracted from diverse linguistic sources. It presents validation mechanisms, employs human judgment within automated frameworks, and envisions collaborative paradigms integrating human expertise with technological algorithms.

.

Keywords: Web data, Skills intelligence, Web intelligence hub, Online job advertisements,

1. Introduction

Skills intelligence is essential for understanding and navigating trends in the dynamic European Union (EU) labour market landscape. Fostering a skilled workforce is not just about acquiring competencies, but strategically aligning them with the evolving demands of industries. Skills intelligence enables individuals and policymakers alike to discern emerging needs, anticipate shifts in demand, and tailor educational and training programs accordingly. By harnessing data-driven insights into skill gaps, job market dynamics, and technological advancements, stakeholders can proactively steer towards a more agile and resilient labour ecosystem. In an era defined by rapid change, the ability to glean intelligence from skills data becomes advantageous and imperative for driving sustainable growth and competitiveness within the EU.

Online job advertisements (OJA) have emerged as a vital data source to provide timely and granular information for skills intelligence, enriching our understanding of labour market dynamics. Analysing these postings unveils valuable insights into evolving skill requirements,

emerging job roles, and industry trends. For more than a decade, Cedefop has been working on extending its skills intelligence toolkit with a data production system (DPS) for gathering and classifying OJA across Europe[1] (more details of the data production system are described in Cedefop (2019).

Cedefop collaborates with Eurostat's Web Intelligence hub (WIH) to examine the possibility of using OJA to develop official statistics on skills. This collaboration underscores a shared commitment to providing accurate and comprehensive data that policymakers, researchers, and practitioners can rely on. By pooling resources, expertise, and methodologies, Cedefop and Eurostat have worked towards harmonising data collection frameworks, standardising skill indicators, and enhancing the quality and reliability of skills-related statistics (Descy et al. l (2019)). By forging closer ties between skills intelligence and official statistics, this collaboration strengthens the evidence base for policy formulation and evaluation, ultimately fostering a more responsive and effective skills ecosystem within the EU.

Although Cedefop already uses OJA data produced by WIH DPS for analytical work (e.g. Napierala (2023) or Cedefop (2023)) the data are still subject to testing for consistency and cross-country comparability. Mainly, as all OJAs are processed in their original language, the ability to extract the skills concepts is under permanent scrutiny. This article describes activities to introduce consistent and cross-language comparable extraction of skills terms. The next part describes automatised methods deployed to extract skills concepts in OJAs and presents the challenges and issues of cross-country comparative results. The third part describes how human judgment and expertise helped us to enhance the precision and uniformity of data extracted from diverse linguistic sources. This part also describes the results of this process. The final part concludes on deploying human judgment within automated frameworks and ideas for further development of this work.

2. Extraction of skills concept in WIH DPS

The extraction of skills concepts (skill classification) is based on ontology matching techniques. For this purpose, ESCO[2] is used as an overarching taxonomy which identifies and categorises skills, competences, qualifications and occupations relevant for the EU labour market and education and training across all EU languages. ESCO skills pillar (version 1.1.1 used in WIH

[1] The data production system covers EU27,

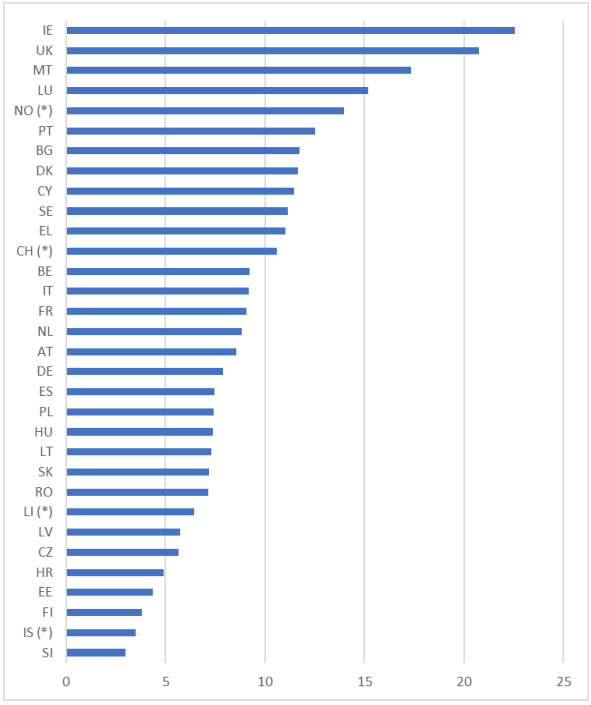
[2] ESCO is the multilingual classification of European Skills, Competences, and Occupations. ESCO is part of the Europe 2020 strategy (for more information, see https://esco.ec.europa.eu/en)

DPS) currently encompasses more than 14,000 concepts (terms), which are broken down to skills, knowledge, transversal skills and languages. Each concept comes with one preferred term and several non-preferred (alternative) terms in each of the ESCO languages.

The process of skill classification is iterative. It begins with the preferred and alternative labels of each ESCO skill being processed through a word-embedding model to derive the most closely related n-grams, which serve as the skill's indicators. These indicators are employed to scan and match skill-relevant text within OJA descriptions. The associated skill is attributed to the OJA whenever an indicator is detected. This matching system operates on a many-to-many basis, where multiple indicators may refer to a single skill, and conversely, one indicator may signal multiple skills.

Despite using the multilingual taxonomy, extracting skills from OJAs presents considerable complexities. In contrast to job titles or other "structured" variables, which tend to be explicitly mentioned or present in pre-defined fields, skills requirements are usually expressed in the unstructured "free text" sections of OJA. Thus, skills descriptions are often influenced by linguistic nuances and the diversity of terminology (use of professionalism or jargon), which may not be fully compatible with the underlying taxonomy. As a result, the DPS can track only a small part of the ESCO across the analysed OJAs.

Figure 1: Average number of skills concepts extracted in one OJA



Source: WIH-OJA data monitoring system (2023q4), note: * country with data only for 2023

The highest "skills yield" per one OJA is achieved in Ireland and the UK (and to some extent Malta). The main reason is that English is the predominant language in these countries' job postings, while English is the "working language" of ESCO[3]. The other reason is the relative simplicity of English grammar in terms of declensions[4] or the fact that most of the algorithms used are developed primarily to "work in English."

3. Securing more even skills extraction

The essential element in developing cross-country comparable official statistics on skills is to test the ability of the data production system to generate comparable and robust yields across all languages. Therefore, the testing and improvements of the DPS aim at two essential tasks to enrich and unify ontologies by:

a) Extending alternative labels of skills concepts across all languages into existing ontology,
b) Adding new and emerging skills concepts which existing ontology does not (yet) contain.

3.1 Extending alternative labels of skills concepts across all languages into an existing ontology

Being aware of differences in occupational structure, linguistic peculiarities, and cultural dimensions of phrasing OJAs, the central assumption of this approach is that OJAs are composed similarly in terms of style and length. This is why we expect WIH DPS to extract a relatively comparable tally of skills for each country and occupation. In other words, the assumption is based on the fact that in the comparative occupation structure across countries, the skills identified in English are also supposed to be identified in other languages. Therefore, the skills augmentation process looked at the most frequent 1,000 skills present in the English pipeline. Overall, the WIH DPS tracks over 2,400 skills in English, but this list's "long tail" features hundreds of skills with only a few mentions, which is unlikely to be found in much smaller language datasets.

[3] ESCO skills terms are primarily developed in English and translated to other languages afterwards. This may cause some losses in translation.

[4] For example, Slavic languages are characteristic of seven cases for each singular and plural in each gender (feminine, mascule, neutral).

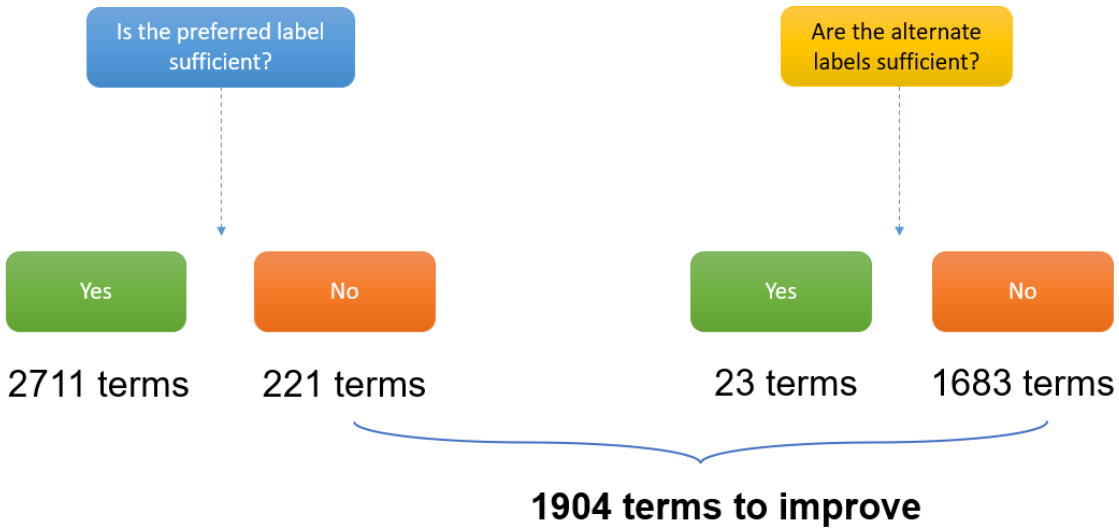
The main phase of this process involved the team of WIH-OJA country experts (ICEs) who were given a list of skills identified in English and not in the country's language. Experts were asked to review the list of terms and answer two questions:

1 - Is the ESCO skill term preferred label sufficient?

2 – Are the alternate labels as described in ESCO sufficient?

In the case of a "no" response to any of these two questions, the experts were asked to identify an alternative term/phrase that would be more suitable. The summative responses are presented in Figure 2.

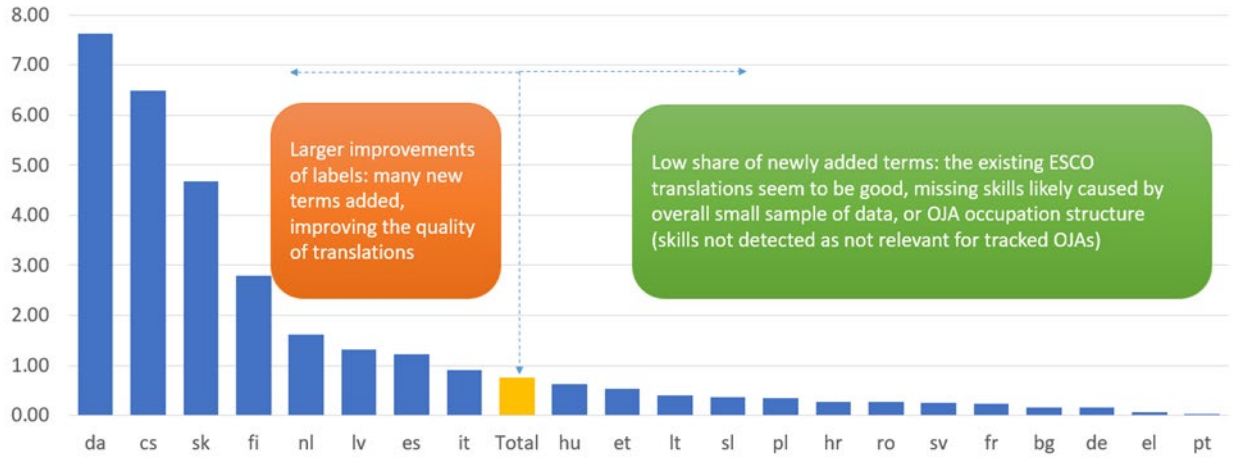
Figure 2: Summary indication of responses to leading questions



Source: ICEs responses

According to scrutinised results, the ICEs' assessment of the suitability and quality of the "preferred labels" was generally considered good. In less than 10% of cases (221 out of 2932) the preferred labels were considered insufficient, and alternative terms were proposed. However, in the alternative labels were considered by ICEs as insufficient, Alternate labels and improvements were proposed in almost 1,700 cases. In total, ICEs proposed to improve about 1.9 thousand terms in the current ontology. As in many cases, ICEs identified more than one new alternate label, and about 3.8 thousand language-skill combinations were proposed to be included in the WIH DPS ontology.

Figure 3: Percentage share of proposed new terms



Source: ICEs responses

As indicated in Figure 3, which represents the percentage share of the proposed corrections on already detected skills in a language pipeline, the most noticeable improvements were proposed for the Danish, Czech and Slovak language pipeline. However, in many languages, the share of newly proposed terms on the total number of already detected skills remained below 1 per cent. The missing skills can result from either an overall small sample of data or differences in OJA occupation structure. The other option is based on the assumption that skills not detected are unimportant for employers and thus not mentioned in OJA. We will follow up with further investigation to clarify this issue.

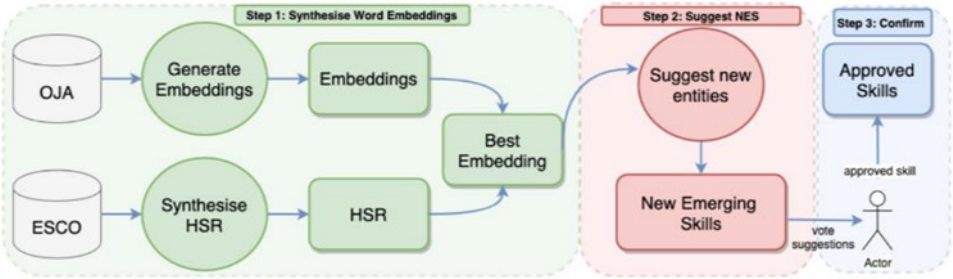
3.2 Adding new and emerging skills concepts which existing ontology does not (yet) contain.

Adding new and emerging skills into an ontology is based on the assumption that the “new” skills are usually presented in the part of the OJA where we can identify skills which already exist in our ontologies. Only that they are “invisible”, in the sense that they are not detected by the ESCO matching algorithm. This happens for a variety of reasons, but likely two of them prevail: new skills (especially in the digital domain), appear frequently, while ESCO is usually

updated bi-annually; employers use different terms than those existing in ESCO. Therefore, we base our work on the Skills classifier and identify n-grams, which often occur repeatedly in the dataset. At this stage, it is unclear if we have captured a new skill, alternative label for existing skills, “spillover” skill from other occupations, or simply an error.

The system is based on word embeddings - a type of word representation that allows words with similar meanings to an equal representation. The key idea behind word embeddings is that words with similar frequency distributions tend to have similar purposes. Words are represented by semantic vectors, which are usually induced from a large corpus using co-occurrence statistics or neural network training. Word embeddings can capture the context of a word in a document and the semantic and syntactic similarity between words and other linguistic patterns.

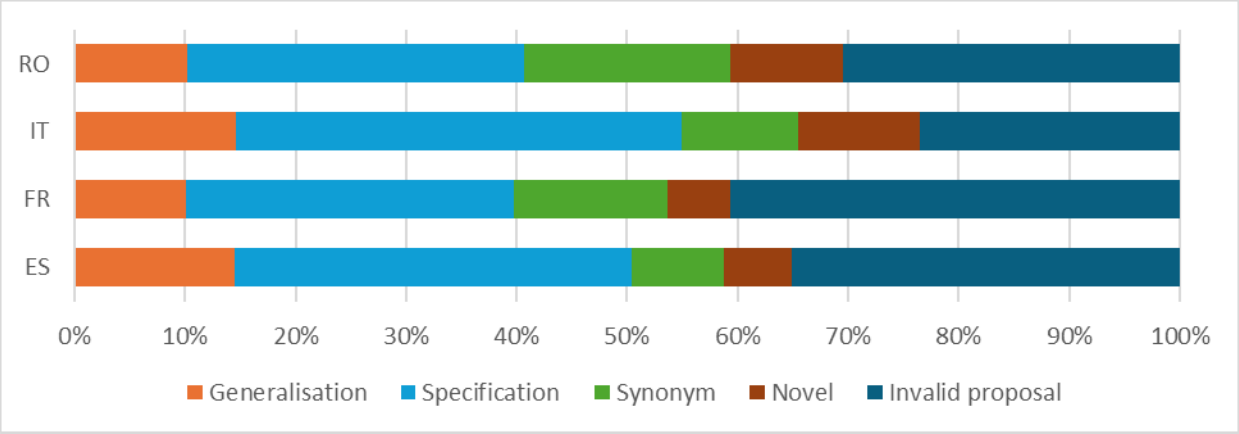
Figure 4: Word embeddings process



Note: visualisation was developed as a part of internal deliverables

The performance of this process was initially pre-piloted in the English language. After a successful pre-pilot, the work was extended to Italian, Spanish, French and Romanian languages. The expressions selected by the system were afterwards shared within the group of experts together with information of occupation that featured the term as well as relevant ESCO skills term in proximity. As the outcome over 800 potential “new” skill concepts in the four languages were identified. After assessment of the terms by the group of human validators the success rate of the tool between 60-80 per cent. Of the valid proposals, largest share was identified as having a relation to an already existing term: either its generalisation, specification, or a synonym.

Figure 5: Assessment of “new” skills terms



Source: ICEs responses

4. Conclusions and next steps

Cedefop and Eurostat have joined forces and throughout Web Intelligence Hub are developing statistics on skills based on OJAs. Although the Data production system is already producing some experimental results, the quality of this data is still under extensive scrutiny. An impotant element of quality control measures is assessment of performance of the of skills extraction across countries, languages and occupations. This ongoing activity uncovered severe disparities across the countries. The utmonst effort is given to understand this phenomena an try to find the ways for improvements.

This article presented two approaches taken to improve skills extractions across languages.The first approchas focused on establishing an unified and cross country comparable list of skills terms and “fill holes” in ontology. The second tried to extract terms used in OJAs which could eventually represent new skills term. Despite of that both approaches strongly built on data-driven techniques and AI the human intervention was of key importance to assess the quality of final outputs.

The next steps of this activity will be to feed this terms back to DPS and assess how it imporvied the overall performance of system. Allongside of this activity Cedefop focusses on the extraction of “specific” skills types such as digital skills or “green” skills (however, more details of this work is subject of other publications). There are also ongoing discussion with ESCO team how to set up a regular exchange of information and contribute to development of the consistent taxonomy of skills. Recently, the work on identification of new skills is scaled up to deliver results for all languages of DPS. Moreover the application of large language models is

tested to provide additional information to new skills terms identified by the AI. However, this activity aims primarily on supporting rather than replacing human intervention. Therefore an effort needs to be made to ensure sufficient access to adequate domain exprets across Europe to contribute to assessment of results.

Acknowledgment

This article buid on results and deliverable of Cedefops contract 2020-FWC7-AO-DSL-VKVET-JBRAN-WIH-OJA/002/20 “Towards the European Web Intelligence hub – European system for collection and analysis of Online Job Advertisements. wich is co-steared by Cedefop and Eurostat via Web intelligence hub.

References (12pt. bold)

Cedefop (2018), Online job vacancies and skills analysis: a Cedefop pan-European approach. Luxembourg: Publications Office,. http://data.europa.eu/doi/10.2801/097022

Descy, Pascaline et al. (2019), Towards a Shared Infrastructure for Online Job Advertisement Data. Statistical Journal of the IAOS, vol. 35, no. 4, pp. 669-675, 2019. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190547, DOI: 10.3233/SJI-190547

Napierala, J. (2023). The feasibility of using online job advertisements in analysing unmet EU demand. Luxembourg: Publications Office. Cedefop working paper, No 18. http://data.europa.eu/doi/10.2801/10233

Cedefop (2023). Going digital means skilling for digital: using big data to track emerging digital skill needs. Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/772175