This project has received funding from the EU Erasmus+ Programme under grant agreement 2021-1-EL02-KA220-YOU-000028780

Co-funded by the Erasmus+ Programme of the European Union

BRIDGING THE GAP

An AI-enabled versatile skill matching tool to assist the less privileged

BRIDGING THE GAP DELIVERABLE 2.1

Authors:	MY COMPANY PROJECTS O.E., International Hellenic University
Status:	Final
Due Date:	31/12/2023
Version:	1.0
Dissemination Level:	PU

Disclaimer:

The content of this document was issued within the frame of the BRIDGING THE GAP project and represents the views of the authors only and is his/her sole responsibility. The European Union/ National agency does not accept any responsibility for use that may be made of the information it contains. The project has received funding from the European Union's Erasmus+: KA220-YOU under Grant Agreement 2021-1-EL02-KA220-YOU-000028780. This document and its content are the property of the BRIDGING THE GAP Consortium. All rights relevant to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the BRIDGING THE GAP Consortium or the Partners detriment and are not to be disclosed externally without prior written consent from the BRIDGING THE GAP Partners. Each BRIDGING THE GAP Partner may use this document in conformity with the BRIDGING THE GAP Consortium Grant Agreement provisions.

(*) Dissemination level. -PU: Public, fully open, e.g. web; CO: Confidential, restricted under conditions set out in Model Grant Agreement; CI: Classified, Int = Internal Working Document, information as referred to in Commission Decision 2001/844/EC.

BRIDGING THE GAP Project Profile

Grant Agreement No.:	2021-1-EL02-KA220-YOU-000028780
Acronym:	BRIDGING THE GAP
Title:	Bridging the Gap – An AI-enabled versatile skill matching tool to assist the less privileged
URL:	https://bridgingthegapproject.eu/
Start Date:	14/02/2022
Duration:	24 months

Partners

INTERNATIONAL HELLENIC UNIVERSITY	DIETHNES PANEPISTIMIO ELLADOS (IHU)	Greece
MYCOMPANY	MY COMPANY PROJECTS O.E.	Greece
	UNIVERSITATEA DIN CRAIOVA	Romania
1992	Regional center for vocational training and education to CCI-Blagoevgrad	Bulgaria

Document History

Version	Date	Author (Partner)	Remarks/Changes
0.1	04/12/2023	Kalliopi Kravari (IHU)	Table of Contents
0.2	11/12/2023	Dimitrios Sarafis (My Company Projects O.E.)	1st Draft ready for internal review
0.3	28/12/2023	Kalliopi Kravari (IHU)	2nd Draft ready for quality control
1.0	29/12/2023	Periklis Chatzimisios (IHU)	FINAL VERSION TO BE SUBMITTED

Abbreviations and acronyms

Deliverable	D
Expected Outcomes	EO
International Hellenic University	IHU
Non-governmental organization	NGO
Labour Force Survey	LFS
Neither in employment nor in education or training	NEET
Human Resources	HR
URI	Uniform Resource Identifier
CV	Curriculum Vitae
JV	Job Vacancy
UoL	Unit of Learning

Executive Summary

BRIDGING THE GAP is a 24 month duration project funding from the European Union's Erasmus+: KA220-YOU under Grant Agreement 2021-1-EL02-KA220-YOU-000028780.

The overarching objective of the BRIDGING THE GAP project is to provide a holistic approach beyond a classical skill-matching system to a system that will bridge the gap as to who are the underprivileged and why are they underprivileged and how education and skill improvement will benefit them.

The main purpose of this document is to a report the progress of the BRIDGING THE GAP project during the deliverable 2.1. More specifically, this deliverable reports on Project Results 2 findings that study a solution framework guide for the EU citizen. The current document analyses the development of a solution framework guide for the EU citizen that will help him/her find a job suitable for him/her or a course that can help to find a better job based only on his/her CV.

Table of Contents

List of Figures & Tables

1 Introduction

1.1 Purpose of the document

The purpose of this document is to present the progress of the project during the second implementation phase regarding the implemented research activities as they are reported in Project Results 2 (deliverable 2.1).

1.2 Intended audience

The intended audience of this document consists of the following target groups:

- BRIDGING THE GAP project partners and the Project Officer at the National Agency
- Young people, especially on the Balkan area, that are interested in skill building/matching
- Labour market actors
- Universities, course providers

1.3 Work Package Objective

The current deliverable 2.1, part of Project Results 2, analyses the development of a solution framework guide for the EU citizen, especially the underprivileged young people. More specifically, Project Results 2 are related to tools development and integrations with external systems & semantic intelligent bridging. For this purpose, all related deliverables including the current 2.1 focus on stakeholders, use cases serving them and building simple user interfaces through prototypes. These prototypes serve as proof of concept that CVs, job descriptions and courses can be annotated with semantics from standard specifications like ESCO. Moreover, we focused on technical issues and developed the back-end of our solution, an intelligent module that matches CVs and job, CVs and courses, Jobs and courses. The rules behind it are coded into this module. Towards this direction, we provide to the public a glass box interface explaining the way the match occurred. Moreover, we incorporate an intelligent software component in order to better guide and support at EU level the validation of non-formal and informal learning should be further pursued especially with regards to the validation of learning acquired through digital resources.

1.4 Structure of the document

In chapter 2, this report describes a short summary of the problem statement and the framework overview as defined by the Bridging The Gap project.

In chapter 3, this report presents the theoretical knowledge that is necessary for the interested reader in order to understand the building blocks and the methods used in our framework.

In chapter 4, this report discusses issues related to all the necessary data required by the developed framework, i.e. CVs from multiple candidates.

In chapter 5, this report describes in detail the architecture of the solution framework while providing all the specifications and details of its implementation.

In chapter 6, this report concludes the findings.

2 Problem Statement and Framework Overview

The COVID-19 pandemic revealed how important it is to ensure people's ability to find employment during emergencies, especially the less unprivileged. The pandemic poses a threat to young people's educational and employment prospects, in addition to causing the loss of life and wealth. The term "underprivileged" refers to disadvantaged students, the unemployed, low-paid, and socially excluded (young) people who make up a significant portion of the population and are underprivileged and underdeveloped from every perspective in society. In general, the lack of education that characterizes this class of people makes them, among others, socially disadvantaged. The project will provide benefits to all EU citizens, with a specific focus on the Balkans throughout its implementation period due to the region's challenges, particularly for young people.

The goal of the project is to develop a semantically rich architecture/framework based on established standards to automate or semi-automate the annotation of CVs while providing a goal-oriented agent-based system that will run without human intervention to persuade employers to accept less-skilled young men/women, whereas the second will have a personal assistant that will identify his/her needs and support life-long training and professional development.

In the following deliverables, we will present pilot applications as proof-of-concept tools where candidates can upload their skills to build their professional profiles and then enrich them with additional educational content sourced from the web of data. By connecting concepts from the ESCO taxonomy with abilities on candidate profiles and learning outcomes from open digital resources, this experimental application will showcase the expected feasibility and potential benefits by linking concepts from the ESCO taxonomy with skills on candidate profiles and learning outcomes derived from open digital resources. More specifically, we believe that such an application can significantly support formal, informal, and lifelong learners in obtaining the necessary skills to enhance their certifications.

3 Theoretical knowledge

In order to enable readers to grasp the framework, first of all, we will present the necessary theoretical knowledge. To start with, this framework is built on techniques and methods coming from the field of Natural language Processing (NLP). Natural language processing (NLP) is a field of computer science and artificial intelligence (AI) that aims to enable computers to comprehend and process human language in a way that resembles human understanding. NLP integrates computational linguistics, which utilizes rule-based models of human language, with statistical, machine learning, and deep learning approaches. NLP empowers computers to interpret and analyze natural language, replicating the human ability to understand and respond to spoken and written communication. In other words, NLP seeks to bridge the gap between human language and machine comprehension, enabling computers to grasp the nuances and complexities of human communication.

These technologies, working in conjunction, empower computers to process human language comprehensively in the form of text or voice data and to 'understand' it's full meaning, including the nuances of intention and emotional context expressed through language.

Natural language processing (NLP) plays a pivotal role in various computer applications, enabling them to translate text, facilitate voice-activated interactions, and summarize large amounts of text swiftly, even in real time. NLP's ubiquitous presence has made it commonplace in our daily lives, from voice-operated GPS systems and digital assistants to speech-to-text dictation software and customer service chatbots. Additionally, NLP has emerged as a game-changer in enterprise solutions, streamlining operations, boosting employee productivity, and optimizing critical business processes.

However, the inherent ambiguities of human language pose a significant challenge to NLP systems. Programming accurate software that can decipher the intended meaning of text or voice data remains an arduous endeavor. The complexities of language, encompassing homonyms, homophones, sarcasm, idioms, metaphors, grammar variations, and sentence structures, demand meticulous training to ensure NLP applications can effectively interpret human language.

Several NLP tasks break down human text and voice data, assisting computers in comprehending the information they intake. These tasks include:

- Speech recognition: Also known as speech-to-text, this process converts spoken language into written text. It is crucial for applications that respond to voice commands or answer spoken questions. Speech recognition faces numerous challenges due to the dynamic nature of human speech, such as slurred words, varying intonation, accents, and grammatical errors.
- Part-of-speech tagging: Also referred to as grammatical tagging, this technique determines the grammatical category of a word or phrase based on its usage and context. For instance, the word "make" can be identified as a verb in "I can make a paper plane" and as a noun in "what make of car do you own?"
- Word sense disambiguation (WSD) is a technique that attempts to determine the most appropriate meaning of a word in a particular context, even if the word has multiple possible

meanings. For example, WSD can distinguish between the meanings of the verb "make" in "make the grade" (succeed) vs. "make a bet" (place a wager).

- Named entity recognition (NER) identifies and classifies specific words or phrases as entities, such as locations, organizations, or people. For instance, NER would recognize "Kentucky" as a location and "Fred" as a person's name.
- Coreference resolution (CR) is the task of identifying and linking mentions of the same real-world entity across a text. For example, CR would determine that the pronoun "she" refers to the same person named "Mary" throughout a text. CR can also identify references to metaphorical or idiomatic expressions, such as recognizing that "bear" in the phrase "speak of the devil" refers to a person who has just been mentioned.
- Sentiment analysis, also known as opinion mining, is the task of identifying and analyzing subjective information in text, such as attitudes, emotions, opinions, and evaluations. Sentiment analysis can be used to extract the overall sentiment of a piece of text (positive, negative, or neutral) or to identify specific sentiment expressions, such as sarcasm, irony, or anger.
- Natural language generation (NLG) is the process of creating human-like text output. It involves converting structured information into natural language that is grammatically correct, fluent, and understandable to humans. NLG is used in a variety of applications, such as generating emails, summaries, and creative content.
- Spam detection: Spam detection is a crucial component of email filtering systems. By analyzing text patterns, NLP can effectively identify and filter out spam emails, reducing the time and effort required to manage personal and business email accounts.
- Machine translation: NLP plays a pivotal role in enabling accurate and efficient machine translation. Beyond simply replacing words, machine translation algorithms must preserve the meaning, tone, and nuances of the original text while adapting to the target language. Google Translate is a widely used example of NLP-powered machine translation.
- Virtual agents and chatbots: NLP powers the natural language interactions between humans and virtual assistants like Siri and Alexa, as well as chatbots. Speech recognition technology enables these intelligent systems to interpret voice commands, while NLP facilitates natural and meaningful responses in text form.
- Social media sentiment analysis: NLP has emerged as a powerful tool for analyzing and extracting insights from social media data. By analyzing sentiment expressed in social media posts, reviews, and responses, NLP enables businesses to gauge customer satisfaction, identify trends, and make informed decisions about product development, marketing strategies, and customer service.
- Text summarization: NLP techniques are employed to summarize lengthy texts, extracting key points and providing a concise overview of the original content. This is particularly useful for

indexes, research databases, or busy readers who require a quick grasp of the main ideas. The most advanced text summarization applications utilize semantic reasoning and NLG to generate summaries that are not only concise but also semantically meaningful and contextually relevant.

From the NLP tasks mentioned above, in our framework, we mostly focus on Named Entity Recognition. In short, Named Entity Recognition (NER) is a subtask of natural language processing (NLP) that involves identifying and categorizing named entities in text. Named entities refer to specific entities, such as people, places, organizations, and products, that are mentioned in a text. NER systems use machine learning algorithms to automatically identify and label named entities in text, enabling users to extract relevant information from large volumes of unstructured data. NER has numerous applications in fields such as information retrieval, sentiment analysis, and recommendation systems. NER has become increasingly accurate in recent years due to the development of deep learning models and the availability of large annotated datasets.

A Named Entity Recognition (NER) system is an interesting tool in NLP since it provides identification of proper nouns in open-domain, namely unstructured, text. NER also known as entity extraction is a subtask of information extraction that seeks to locate and classify elements in text into predefined categories like names of persons, organizations, locations, etc. Using NER the necessary attributes from the resume can be extracted by training samples of data. Key role of NER in our study is to find the how the same token was tagged in different parts of the same document.

Similar methods have been used in older and recent works too. For example in the work of Tim Zimmermann et al. [1], they show how to use a data-driven approach including multiple datasources in order to guide a user in matching candidates and jobs. Using ML and NLP, it is possible to build a pipeline that first extracts all the relevant information from resumes and provides them in a structured way. Once resumes are processed, external data for employers and educational institutes are included as well to calculate candidate matches. Recruiters can tailor their search and filtering to specific job roles. Additionally, there are several options and dimensions to compare candidates. Eventually, the application allows for a detailed analysis of the resume to validate ranking and matching recommendations.

Another example is the work of Saket Maheshwary et al. [2] where they investigate the important and challenging task of recommending appropriate jobs for job seeking candidates by matching semi structured resumes of candidates to job descriptions. To perform this task, they propose to use a Siamese adaptation of convolutional neural network. The proposed approach effectively extracts the underlying semantics, enabling it to project similar resumes and job descriptions closer together, and to project dissimilar resumes and job descriptions further apart in the semantic space. The experimental results on a set of 1314 resumes and a set of 3809 job descriptions (5, 005, 026 resume-job description pairs) demonstrate that their approach is better than the current state-of-the-art approaches.

Also in another paper [3], the authors tackle the problem of finding the first full-time, major-related job which is a challenge faced by most college students, specifically, those who have not yet gained much working experience face a considerable challenge when entering the job market, particularly students majoring in Information Technology (IT) and cybersecurity. In the IT and cybersecurity domains, the combination of an evolving technology landscape, intense job competition, and increasing expectations from employers poses significant challenges for recent graduates. Their study shows the initial results of a tool called e-RAP which allows the students to submit their current resumes, obtain automatic feedback and a rating report, and consequently take actions to strengthen their portfolio.

The authors developed a resume analysis and reporting tool by leveraging machine learning and natural language processing (NLP) techniques. The methodology section dives into the e-RAP analysis process, shedding light on data curation, data collection, and data analysis methods employed. E-RAP generates insightful reports that visually demonstrate the tool's ability to enhance student resumes and identify skill areas that require improvement or emphasis. The authors conducted a comprehensive evaluation of over 60 resumes processed through e-RAP, demonstrating the tool's potential to enhance the quality of student resumes.

Furthermore, in this paper [4], a relationship network between skills in recruitment domain by using the neural net inspired by Word2vec model is proposed. It is observed that it is possible to train high quality word vectors using very simple model architectures due to lower cost of computation. Moreover, is seems that we can compute quite accurate high dimensional word vectors when we use much larger datasets. Using Skip-gram architecture and an advanced technique for preprocessing data, the result seems to be impressive. The result of their work can contribute to building the matching system between candidates and job post. In the other hand, candidates can find the gap between the job post requirements and their ability, so they can find the suitable training.

Finally, in this paper [5] the authors explore the feasibility of creating a standard parser for resumes of all formats. They found that this was not possible to do without information loss for all cases, which would result in the unfair loss of certain applicant’s resumes in the process. Instead they proceeded with LinkedIn format resumes, for which they could build a parser with no information loss. This parser is available in a web application. They used BERT for sequence pair classification to rank candidates as per their suitability to a particular job description. With the data collected from the web application, they had real job descriptions and interviewer comments at each stage of the hiring process. Their work establishes a strong baseline and a proof of concept which can lead to the hiring process benefiting from the advances in deep learning and language representation.

4 Data: Curricula Vitae

The solution framework for the EU citizen is based on the Curricula Vitae or CV of each citizen. A Curricula Vitae (CV) is, typically, a document that summarizes a person's academic and professional history, accomplishments, skills, and qualifications. It is mostly used to apply for academic positions, research opportunities, and jobs in many industries. A well-prepared CV provides a snapshot of a person's education, work experience, publications, awards, and other relevant information, and is often tailored to a specific job or field. The goal of a CV is to convince a potential employer or institution that the applicant is a good fit for the position and has the necessary qualifications and expertise to excel in the role.

It usually consists of the following sections:

- Personal information: This includes full name, contact information, and other personal details such as date of birth, nationality, and marital status.
- Education: This section lists academic qualifications, including degrees, diplomas, certificates, and any other relevant training or courses the candidate has completed.
- Work experience: This section provides a detailed summary of work history, including the positions the candidate had held, the companies worked for, and the dates of employment. It should also include a brief description of job responsibilities, achievements, and any notable accomplishments during each role.
- Skills: This section highlights technical, transferable, and soft skills, including language proficiency, computer literacy, communication skills, and problem-solving abilities.
- Publications and presentations: If the candidate authored or co-authored any publications or presentations, this section provides a list of the titles, dates, and relevant details.
- Awards and honors: This section lists any awards, scholarships, or honors received in recognition of achievements.
- Professional memberships: This section lists any professional organizations or associations the candidate belong to, including the dates of membership and any leadership roles he/she has held.
- References: This section provides the names and contact information of individuals who can vouch for the candidate’s work experience, skills, and qualifications.

From all the sections mentioned above, the only bullets necessary for our framework’s functionality are the skills and the working experience of the candidate. The reason for that is the structure of the ESCO framework which uses Skills and Occupations as the common vocabulary between jobs, courses and CVs to allow matches to happen between them.

Details on the acquisition of the CVs for our database can be found in deliverable 2.4.

5 Solution Framework

This section presents the architecture of the solution framework that was built in the framework of this project while discusses the implementation details of it.

5.1 Architecture

The implementation of the solution framework consists of several modules as presented in the list and the accompanying figure below. In the following list, a short description of each module is given. Information flows through the modules in the order they are listed below.

1. **Text Extraction Module:** The first module serves as the starting point for the processing workflow. It extracts text from a text file and performs any necessary pre-processing tasks such as converting all text to lowercase and tokenizing the text.
2. **CVs Database**: All CVs stored in the database has undergone annotator review prior to insertion in the database, enabling us to extract relevant rules.
3. **Rules and Filters**: This module meticulously examines the CVs stored in the database to extract valuable rules and exceptions, which will be utilized as filters during the annotation process.
4. **Annotator Module**: In this module a variety of different processes take place as the annotator is the main component of the solution framework. The annotator's central task in this module is to extract skills and occupations from the candidate's CV employing standard NLP methods and the ESCO terminology.
5. **Post Processing Module**: In the final step of the implementation, the extracted skills and occupations undergo organization and refinement (including removing duplicates) before being presented to the user.

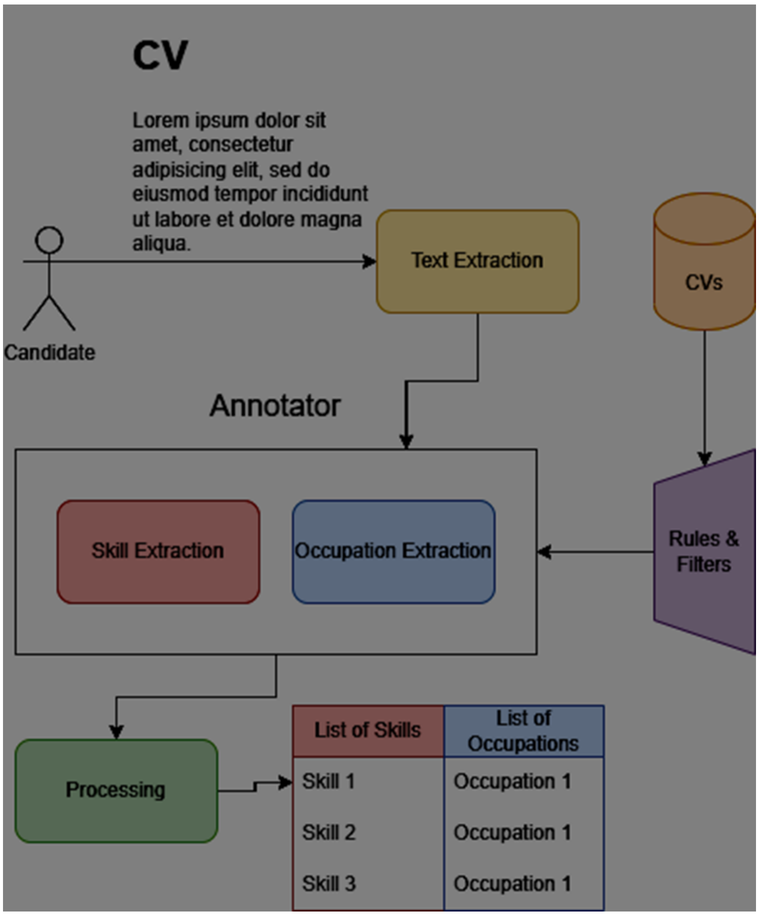

Figure 1: Architecture of Solution Framework.

Before delving into the detailed description of our algorithm, we'll provide a brief overview of the concepts underpinning it to ensure clarity for the reader. At its core, our algorithm hinges on a function that identifies common skills and occupations across all CVs in our dataset. To effectively connect CVs with multilingual and multi-sector Jobs and Courses, we employ the ESCO framework, which establishes a universal vocabulary of skills and occupations, enabling us to seamlessly link all entities within our data (CVs, Jobs, Courses). In the following paragraphs, we will elucidate the ESCO framework and its key concepts essential for comprehending our algorithm.

ESCO (European Skills, Competences, Qualifications, and Occupations) serves as the European multilingual classification system for Skills, Competences, and Occupations. It functions as a comprehensive dictionary, effectively describing, identifying, and classifying relevant professional occupations and skills within the EU labor market and education and training sector. These concepts and their intricate relationships are interpretable by electronic systems, enabling various online platforms to utilize ESCO for diverse services. These services include matching job seekers with suitable employment opportunities based on their skills, suggesting training programs for individuals seeking to enhance or acquire new skills, and so on.

The ESCO dictionary consists of 3,008 occupations and 13,890 skills associated with these occupations, translated into 28 languages, including all official EU languages, along with Icelandic, Norwegian, Ukrainian, and Arabic. ESCO's main objective is to facilitate job mobility across Europe, and therefore contribute to a more integrated and efficient labor market. To achieve this goal, ESCO offers a unifying "common language" for occupations and skills that can be used by various stakeholders involved in employment and education and training matters.

The skills pillar, one of the three pillars of ESCO, adopts a broad definition of skills, encompassing not just skills but also knowledge and competences. Within the skills pillar, ESCO distinguishes between two types of skill concepts: skill/competence concepts and knowledge concepts. This distinction is made by indicating the skill type. However, there is no differentiation between skills and competences within the recorded ESCO skills pillar.

ESCO v1 comprises approximately 13.500 knowledge, skills, and competence concepts. Each of these concepts is associated with one preferred term and an unlimited number of non-preferred terms and hidden terms in each of the ESCO languages. Additionally, each concept includes an explanation in the form of a description, scope note, and definition. The following list provides a summary of the metadata for ESCO knowledge, skill, and competence concepts and their connections to other ESCO pillars:

- **Preferred term**
- **Non-preferred terms**
- Hidden terms
- Description
- Formal definition
- Scope note
- Skill type (knowledge or skill/competence)
- Skill reusability level
- **Occupations for which the knowledge, skill or competence is essential**
- **Occupations for which the knowledge, skill or competence is optional**
- **URI**

In 2019, following discussions with the ESCO Member States Working Group and the ESCO Maintenance Committee, the Commission, with the support of a group of experts, started the development of a hierarchy for the 13.485 skills and knowledge concepts of ESCO. The skill and knowledge hierarchy aim to enable users to search and retrieve the ESCO skill and knowledge concepts systematically for a variety of purposes and support the matching of jobseekers with job vacancies. The ESCO skills hierarchy can also facilitate the mapping exercise foreseen in Commission Implementing Decision (EU) 2018/1020 on the adoption and updating of the list of skills, competences and occupations of the European classification for the purpose of automated matching through the EURES common IT platform.

The ESCO skills and knowledge hierarchy is a single all-embracing hierarchical framework containing four distinct sub-classifications:

- Knowledge
- Skills
- Attitudes & values
- Language skills and knowledge

The developers of the ESCO hierarchy mapped the 332 Intermediate Work Activities (IWA) specified in US O*NET to the second level of the hierarchy. Most of the third level categories were created based on mapped IWAs that were used either singly or clustered to form level three categories.

The occupations pillar is the second of the three pillars of ESCO. It organises the occupation concepts. It uses hierarchical relationships between them, metadata as well as mappings to the International Standard Classification of Occupations (ISCO) in order to structure the occupations.

ESCO v1 contains 2.942 occupations. Each occupation concept contains one preferred term and any number of non-preferred terms and hidden terms in each of the ESCO languages. It also includes information on regulated professions that are relevant in the context of this occupation.

Each occupation also comes with an occupational profile. The profiles contain an explanation of the occupation in the form of description, scope note and definition. Furthermore, they list the knowledge, skills and competences that experts considered relevant terminology for this occupation on a European scale. ESCO distinguishes essential and optional knowledge, skills and competences. The following list provides an overview of the metadata for ESCO occupations and relationships to other ESCO pillars:

- **Preferred term**
- **Non-preferred terms**
- Hidden terms
- Description
- Formal definition
- Scope note
- Information on regulation
- ISCO-08 code
- **Essential skills and competences**
- **Optional skills and competences**
- Essential knowledge
- Optional knowledge
- **URI**

The International Standard Classification of Occupations (ISCO-08) serves as the hierarchical structure for the occupations pillar. Each ESCO occupations is assigned to exactly one ISCO-08 unit group (level

4). Thus, ISCO-08 provides the top four levels for the occupations pillar. ESCO provides the fifth and lower levels of the hierarchical structure with its list of occupations. In addition, ESCO contains hierarchical relationships between ESCO occupations. This way, ESCO shows that specialisms are narrower in scope than a more generic occupation. For example, ESCO contains an occupation "bartender" and two specialisms "cocktail bartender" and "barista", which are hierarchically linked to it.

The qualifications pillar is the third of the three pillars of ESCO. Information on qualifications at European level is now displayed in Europass, and comes from databases of national qualifications reflecting the National Qualifications Frameworks that are owned and managed by the EU Member States. Europass offers the most up to date and rich repository of high quality data on qualifications, national qualification frameworks and learning opportunities in Europe, helping learners to find a course in another country and employers to grasp the value of a qualification from a different EU Member State

The qualifications pillar supports the understanding of the individual qualifications needed by employers, public and private employment services, learners, workers, jobseekers, education and training providers and other actors. This information should be as complete and transparent as possible to meet the information needs of these stakeholders. Therefore, only qualifications data which includes the following core information are displayed in Europass, based on Annex VI of the EQF Recommendation:

- Title: Exact title of the qualification (without translation).
- Field: Based on ISCED Fields of Education and Training 2013.
- Country/Region
- European Qualifications Framework (EQF): only relevant for qualifications that already have an EQF level assigned through the referencing process of National Qualifications Frameworks to the EQF.
- Awarding body or competent authority.
- Description of the qualification expressed in learning outcomes

Information on other fields can also be provided, such as credit points, internal quality assurance process, link to qualification supplement, entry requirements, etc.

Member States and other stakeholders wishing to publish information on their qualifications in ESCO need to structure their data according to the qualifications metadata schema developed for this purpose and upload it in the qualifications dataset register (QDR).

A labour market terminology that can help understanding which occupations and skills are related to a particular qualification allows learners, job seekers and employers to best use this information: ESCO fills this need by providing an updated, evidence – based and multilingual skills and occupation vocabulary

Since the publication of ESCO v1.0, organisations providing data on qualifications can link learning outcomes of qualifications with ESCO terminology. To this end, they identify knowledge, skill and competence concepts in the skills pillar of ESCO that are relevant in the context of the learning outcome description of a qualification. In 2019 and in 2020, the European Commission conducted a pilot project with the Member States in order to test automated linking of learning outcomes of qualifications with ESCO skills in different languages and developed a dedicated IT tool to support national authorities in this exercise. The project demonstrated the value of using the ESCO skills thesaurus to provide transparency of qualifications and better quality of data on individual learning outcomes. A second phase of the pilot took place in the second half of 2020.

Relationships between qualifications and the occupations pillar of ESCO are only displayed, if they already exist at the national level. Member States are not developing such data for ESCO. The relationship can indicate e.g. if a qualification is a requirement in order to perform an occupation in the specific Member State.

From the pillar concepts presented above we decided to move on with the skills and occupations and leave qualifications out. The reason for this decision was that these two are the concepts of the ESCO framework that can can be used by us to connect the entities (CV, Jobs, Courses) to create the common vocabulary mentioned in the introduction of this chapter.

As mentioned above, ESCO skills have many attributes, from labels associated with them to short descriptions given to them. They are also organized in a hierarchical structure for them to be more easy to find and to employ in some other kind of framework. From all those attributes mentioned above we decided to use the following for the Skill concept:

- Preferred term
- Non-preferred terms
- Occupations for which the knowledge, skill or competence is essential
- Occupations for which the knowledge, skill or competence is optional
- URI (Uniform Resource Identifier)
- and the following for the Occupation concept:
- Preferred term
- Non-preferred terms
- Essential skills and competences
- Optional skills and competences
- URI (Uniform Resource Identifier)

The reasons behind this decision is the fact that these attributes are sufficient for our framework and by extension for the algorithm presented in the next chapter which is the implementation of it. This decision will become match clearer in the next deliverables of this work package.

Apart from the Preferred or Non-preferred term (or label) for which we will give a definition in the following paragraphs, the rest of the attributes are pretty self-explanatory.

Preferred term definition from ESCO: Each concept within ESCO has a designated, unique preferred name per ESCO language. It is called the preferred term (PT) and can be a single-word term or a multi-word term. The preferred term is used to represent a concept in ESCO in a specific language. Out of a group of terms with similar meaning, the one that best represents the concept is chosen to be the preferred term. The preferred term of a given concept is unique per language.

Alternative terms definition from ESCO: Non-preferred terms (NPTs) can be synonyms (words with similar or same meanings) but can also be spelling variants, declensions, abbreviations, etc. They are regularly used by the target group (jobseekers, employers, education institutions) to refer to concepts that are described in ESCO with the preferred term.

5.1 Algorithm description / Implementation

In this section we are going to present one-by-one the steps of the algorithm we built in order to extract ESCO skills and occupations from a given text.

The first step is to load the ESCO vocabulary in our system and keep as mentioned before only the attributes necessary for us which are the preferred and the non-preferred labels. The labels are different for each language in our system, this means that for each language the preferred label of a specific occupation is going to be different although the occupation behind it is going to be the same.

The next step is to set for each language a list of skill and occupation exceptions. We decided to do that because after experimenting for a while with the data, we discovered that some preferred or non-preferred labels, either for skill or occupation, are a very common or generic word of the respective language. For example, there is a skill's label named "energy" which is an ambiguous word that can be interpreted differently depending on the context and also appears quite often in a piece of text. As a result, a skill named "energy" would be assigned to almost every CV without adding any useful information and it also made the algorithms of our system that worked on these data, harder to perform. Hence, we decided to create an exception list for each language, to keep these cases of labels out of the returned output of the algorithm.

After initializing the necessary variables mentioned above, we proceeded with the processing of the input text. The processing steps are presented in the following ordered list for purposes of convenience.

1. The text that is going to be processed is given as input to the algorithm.
2. All the text is converted to lowercase as all labels, either from skills or occupations, are in lowercase. Next, all non-alphanumeric characters are removed from the string.
3. The converted text is passed through a tokenizer from the Python's spaCy library. A different tokenizer is selected for each language. Tokenization is the process of breaking down a large text into smaller pieces, called tokens, such as words, phrases, symbols, or even subwords.
4. We filter out all tokens that fall in the exception list. We also filter those that have number or any other non-vocabulary symbols in them, except for some special characters (. # + * -),

because ESCO labels are almost purely alphabetical and because we need to reduce the number of tokens in order to make the searching process faster.

5. From the tokens generated in the previous steps, we create sequences of length 2 or 3. This is a necessary step as many of ESCO labels are consisted of two or three words.
6. After that, we start searching for matches between skills' labels and tokens or sequences. For each match found, we add the corresponding skill to the output list.
7. We repeat the same process for the ESCO occupations.
8. Finally, we remove any label duplicates and we return the output lists.

The libraries and their versions used in this algorithm's implementation are listed in the following table.

Table 1: Libraries and their Versions used in the framework's implementation

Library	**Version**
Python	3.8
nltk	2.9.0
spaCy	3.3.1
pdfminer-six	2.1.3
pandas	1.4.3

6 Conclusion

This deliverable provided a detailed description of the development process for the solution framework for the EU citizen. More specifically, the problem of matching EU citizens with suitable jobs across the EU is a complex one. The EU has a diverse workforce with a wide range of skills and experiences, and there is a need for a system that can effectively match these individuals with the right job opportunities. Our solution framework aims to address this challenge by providing a tool that can identify and connect EU citizens with potential job openings. Our framework is built on a foundation of natural language processing (NLP) and machine learning (ML) techniques. NLP is used to extract relevant information from CVs, such as skills, experience, and education. ML is used to identify patterns in the data and to generate recommendations for matching EU citizens with potential job openings. The key data input for our framework is CVs from multiple EU citizens.

Our solution framework is composed of four main components:

Data Preprocessing: This component cleans and prepares the CV data for analysis. This includes tasks such as removing HTML markup, converting text to lowercase, and tokenizing the text.

Skill and Experience Extraction: This component extracts relevant skills and experience from the CV data. This is done using NLP techniques such as part-of-speech tagging and named entity recognition.

Skill and Occupation Matching: This component matches the extracted skills and experience with a database of job openings.

Recommendation Generation: This component generates recommendations for matching EU citizens with potential job openings. This is based on the matching results from the Skill and Occupation Matching component.

Our framework is implemented using a Python-based NLP and ML toolkit called spaCy. spaCy provides a range of functionalities for NLP tasks, such as tokenization, part-of-speech tagging, and named entity recognition. For ML tasks, we use scikit-learn, which is a popular Python library for machine learning.

We have evaluated our framework using a dataset of CVs from multiple EU citizens. The results of our evaluation show that our framework is able to accurately match EU citizens with potential job openings. We are continuing to improve our framework by adding new features and refining our algorithms.

Our solution framework has the potential to make a significant contribution to the problem of matching EU citizens with suitable jobs across the EU. By providing a tool that can effectively connect these individuals with the right job opportunities, we can help to improve the efficiency of the labor market and to enhance the lives of EU citizens.

References

[1] Tim Zimmermann, Leo Kotschenreuther, & Karsten Schmidt. (2016). Data-driven HR - Resumé Analysis Based on Natural Language Processing and Machine Learning.

[2] Saket Maheshwary and Hemant Misra. 2018. Matching Resumes to Jobs via Deep Siamese Network. In Companion Proceedings of the The Web Conference 2018 (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 87–88. https://doi.org/10.1145/3184558.3186942

[3] Green, Nathan; Liu, Michelle; Murphy, Diane Information Systems Education Journal, v18 n3 p28-37 Jun 2020

[4] Le Van-Duyet, Vo Minh Quan, & Dang Quang An. (2019). Skill2vec: Machine Learning Approach for Determining the Relevant Skills from Job Description.

[5] Vedant Bhatia, Prateek Rawat, Ajit Kumar, & Rajiv Ratn Shah. (2019). End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT.