



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 2.3

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://bridgingthegaproject.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
	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.)	1 st Draft ready for internal review
0.3	28/12/2023	Kalliopi Kravari (IHU)	2 nd 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 report the progress of the BRIDGING THE GAP project during the deliverable 2.3. More specifically, this deliverable reports on Project Results 2 findings that study a solution framework guide for the EU education institutes that will help them create courses more suitable for candidates to join them and improve their chances in finding a job suitable for them, by incorporating them in our common framework.

Table of Contents

1	Introduction	8
1.1	Purpose of the document.....	8
1.2	Intended audience.....	8
1.3	Work Package Objective.....	8
1.4	Structure of the document.....	8
2	Problem Statement and Framework Overview	9
3	Data: Units of Learning	10
4	Solution Framework	12
4.2	Architecture.....	12
4.2	Implementation.....	13
6.	Conclusion	17
	References	18

List of Figures & Tables

Figure 1: Architecture of Solution Framework. 13

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 2.3).

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 2.3, part of Project Results 2, analyses the development of a solution framework guide for the EU education institutes that will help them create courses more suitable for candidates to join them and improve their chances in finding a job suitable for them, by incorporating them in our common framework. 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 2.3 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 discusses issues related to all the necessary data required by the developed framework, i.e. courses and seminars scraped from multiple course providing websites.

In chapter 4, this report describes in detail the architecture of the solution framework while providing all the specifications and details of its implementation.

In chapter 5, 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 privileged. 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 goal is to enable underprivileged young people the opportunity to look for employment within the EU and develop their talents by making the tools available that result from the project tools to better explain their skills and qualifications with particular, accepted semantics. On the other hand, EU businesses will receive assistance in describing professions and particular employment within a semantic context that has been established. In addition, organizations that provide on-the-job training must adjust their programs and learning outcomes to meet the needs of EU citizens and workers.

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 job descriptions and UoLs 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 Data: Units of Learning

The solution framework for the EU course provider deals with the Units of Learning, or UoLs, of each institute or generally of a course provider.

Online courses refer to educational programs or classes that are delivered through the web, allowing learners to access and engage with course materials remotely using a computer, tablet, or mobile device. These courses are designed to provide learning opportunities outside of traditional classroom settings, enabling individuals to acquire knowledge, develop skills, and earn certifications or qualifications from the comfort of their own location and at their own pace.

Online courses typically feature the following components:

- **Course Content:** Online courses present structured learning materials such as video lectures, reading materials, quizzes, assignments, and interactive multimedia content. These materials are usually organized into modules or lessons that cover specific topics or learning objectives.
- **Learning Management System (LMS):** Online courses are often hosted on a dedicated platform called a Learning Management System. The LMS serves as a central hub where learners can access course materials, participate in discussions, submit assignments, track their progress, and communicate with instructors and fellow students.
- **Flexibility and Self-Paced Learning:** One of the key advantages of online courses is the flexibility they offer. Learners can access the course materials and complete the assignments according to their own schedules, allowing them to balance their studies with work, personal commitments, or other responsibilities. Some online courses also provide self-paced learning options, enabling students to progress through the course at their preferred speed.
- **Instructor Support:** Although online courses are typically self-directed, they often include support from instructors or subject matter experts. Instructors may provide guidance through online forums, discussion boards, email, or virtual office hours to answer questions, provide feedback on assignments, and facilitate discussions among learners.
- **Assessment and Certification:** Online courses often include assessments or quizzes to evaluate learners' understanding of the material covered. Successful completion of these assessments may lead to the awarding of certificates or digital badges, which serve as evidence of the learner's achievement and can be shared on resumes or professional profiles.

Online courses cater to a wide range of subjects, from academic disciplines like math, science, and literature, to professional skills such as programming, digital marketing, or project management. They are offered by universities, colleges, educational platforms, and organizations worldwide, making education more accessible to individuals regardless of their geographical location or time constraints.

Overall, online courses provide a flexible and convenient way to learn and acquire new knowledge or skills, making education more accessible, adaptable, and personalized to individual learners' needs.

From all the sections mentioned above, the only bullets necessary for our framework's functionality are the course content. The reason behind that is the structure of the ESCO vocabulary, which uses Skills and Occupations as the common vocabulary between jobs, courses and CVs, in order for the matches to happen between them.

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

4 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.

4.2 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. **Web Crawler:** This module is responsible for scraping websites containing courses and seminars and extracting any information from them that are useful to our framework.
2. **UoLs Database:** All UoLs stored in the database has undergone annotator review prior to insertion in the database, enabling us to extract relevant rules.
3. **Text Extraction Module:** This module is responsible for extracting text from a text file and performs any necessary pre-processing tasks such as converting all text to lowercase and tokenizing the text.
4. **Rules and Filters:** This module meticulously examines the UoLs stored in the database to extract valuable rules and exceptions, which will be utilized as filters during the annotation process.
5. **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 UoL text employing standard NLP methods and the ESCO terminology.
6. **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.

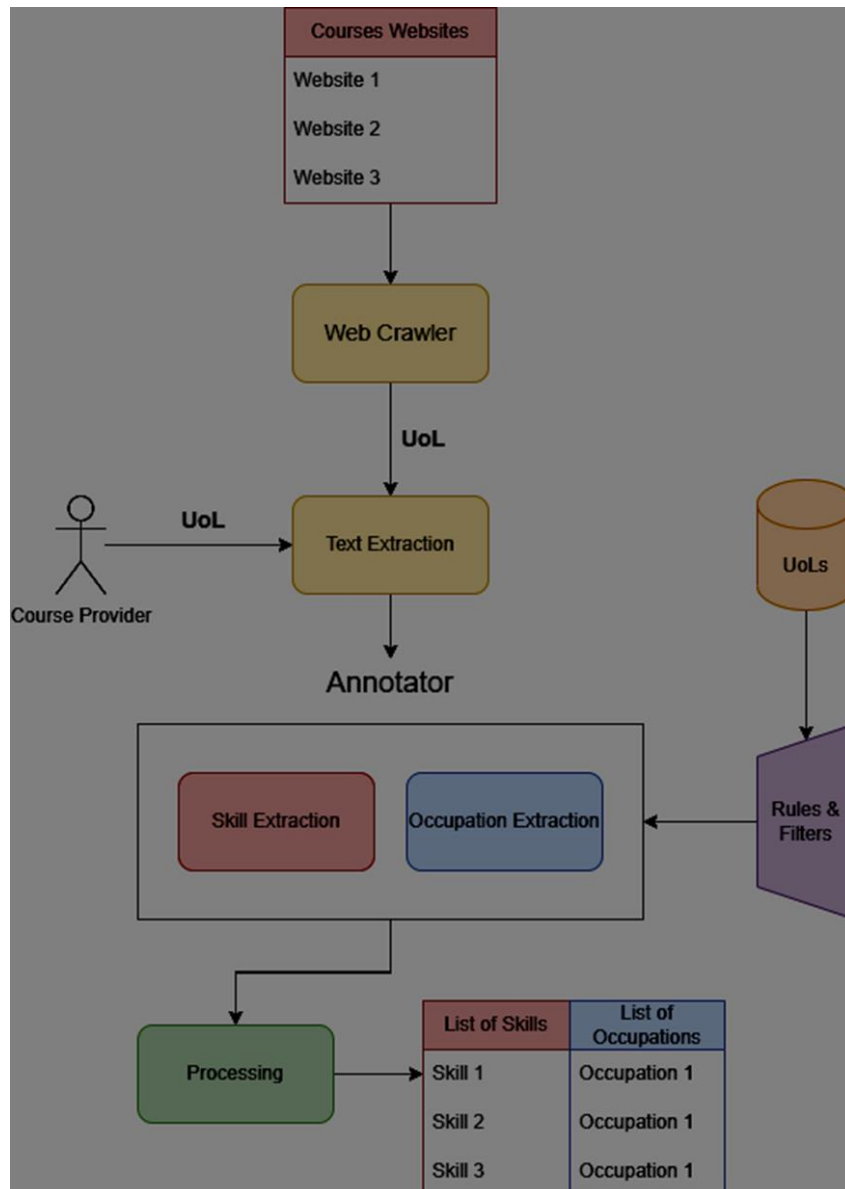


Figure 1: Architecture of Solution Framework.

4.2 Implementation

Prior to delving into the intricacies of the framework's primary function, which involves processing each UoL and extracting relevant Skills and Occupations, we provide a comprehensive description of the Web Crawler module. The position of this module within the framework's architecture is illustrated in figure 1. The primary objective of the Web Crawler module is to populate the database with courses and seminars, ensuring that:

1. We could be able to extract rules from them and use them as filters in the annotator and as a result improve its performance.
2. There are already available UoLs in our system's database so that the candidate can search for courses and courses can be proposed to him/her by our system, regardless of the UoLs the institutes have inserted into the database.

We will now proceed with outlining the sequential steps of the algorithm employed to extract data from the websites. This algorithm adheres to a consistent structure across all the websites. Initially, our first task involved locating a link to the webpage displaying the available job positions on each website. Subsequently, we conducted web scraping on the page referred to by the link, aiming to retrieve a comprehensive list of course pages containing all the essential information we required. After having the content of each course/seminar we scraped that content and we processed it in order to get the following info:

- Title: This feature's value is the title of the course/seminar as scraped from its webpage.
- Provider: This feature's value is the name of the institute/course provider that provides the specific course.
- Link: This feature's value is the URL of the course's webpage.
- Language: This feature's value is the language used in the course's description.
- Description: This feature's value is the course's description as found in its webpage.

Subsequently, the gathered UoL data was sorted and categorized into CSV files based on the source website and the language of the text. Each entry within these files represented an individual UoL, accompanied by the aforementioned attributes. To ensure consistency, the data within each file underwent a data cleaning process and was transformed into a standardized format suitable for input into the annotation algorithm. The following steps outline the methodology employed to process each file:

1. Drop any listings that didn't contain any scraped description because this feature is necessary for the skill and occupation extraction module.
2. Clean the text in each field from any unnecessary characters and make it more uniformly formatted in order to make the extraction algorithm perform faster and improve its robustness.
3. Finally, if a file contained listings of multiple languages, we created separate files one for each language found in the initial file.

The data contained in those files were finally inserted into our database. The process and methods used for inserting those data are going to be described in a later deliverable.

We will now delve into the specifics of the framework's primary module, known as the annotator module. In the following paragraphs, we will sequentially present the individual steps of the algorithm we developed to extract ESCO skills and occupations from a provided UoL text.

To initiate the process, we begin by loading the ESCO dictionary into our system and, as previously stated, selectively retaining the necessary attributes, specifically the preferred and non-preferred terms. It is important to note that each language within our system has its own nomenclature. Consequently, the preferred term for a specific occupation will vary across languages, despite referring to the same underlying occupation.

Subsequently, we proceed to establish a list of skill and occupation exceptions for each language. This decision stems from our extensive experimentation with the data, wherein we discovered that certain preferred and non-preferred terms, whether they pertain to skills or occupations, are excessively common or ordinary words within their respective languages. To illustrate, consider the term "power," which is an ambiguous term that can assume different meanings depending on the context, and it frequently appears in the text. Consequently, the term "power" is assigned as a skill to almost every

UoL, contributing little valuable information. Moreover, this abundance of occurrences poses challenges for our system's algorithms in effectively processing the data. Hence, we opted to create an exception list for each language, excluding such labels from the algorithm's generated output.

Once the necessary variables are initialized, we proceed with the processing of the input text. The following steps outline the sequential order of the processing:

1. The input text is provided as input to the algorithm.
2. To ensure consistency, all texts are converted to lowercase since all tags, whether skills or professions, are in lowercase. Additionally, any non-alphanumeric characters are removed from the string.
3. The transformed text is then tokenized using the appropriate tokenizer from Python's spaCy library. Different tokenizers are chosen for each language. Tokenization involves breaking down the text into smaller units called tokens, which can include words, phrases, symbols, or subwords.
4. Tokens that fall within the exception list are filtered out. Additionally, tokens containing numbers or non-lexical symbols, with the exception of specific characters like (. # + * -), are also filtered. By reducing the number of tokens, particularly non-alphabetic ones, we enhance the efficiency of the search process.
5. Sequences of length 2 or 3 are generated from the tokens obtained in the previous steps. This step is crucial as most ESCO labels consist of two or three words.
6. Subsequently, the algorithm searches for matches between the labels of the skills and the tokens or sequences. For each discovered match, the corresponding skill is added to the output list.
7. The same process is repeated for ESCO occupations, searching for matches between their labels and the tokens or sequences.
8. Finally, any duplicate labels are eliminated, and the output lists are generated.

By following these processing steps, we are able to extract the relevant skills and occupations from the input text, ensuring the output lists are free from duplicates.

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
selenium	4.10
beautifulsoup4	4.12.2
urllib	2.0.3

6. Conclusion

This deliverable provided a detailed description of the development process for the solution framework for the EU course provider. More specifically, we provided a short summary of the problem's definition and the goal of the framework while discussing details about all the necessary data for our framework to operate, which are UoLs scraped from multiple course providing websites. Finally, we presented the architecture of the solution framework and provided all the specifications and details of its implementation.

References

- [1] BRIDGING THE TRANSITION BETWEEN EDUCATION AND THE LABOUR MARKET, Report of the conference of the European Network of Education Councils, Prague, 20-21 October 2014.