This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

REBUILD
ICT-enabled integration facilitator and life rebuilding guidance
*Project start date: 01/01/2019 | Duration: 36 months*

# Deliverable: D3.2 Methodology for skills and needs matching

DUE DATE OF THE DELIVERABLE: 30-03-2020
ACTUAL SUBMISSION DATE: 23-04-2020

| | |
|---|---|
| **Project** | REBUILD – ICT-enabled integration facilitator and life rebuilding guidance |
| **Call ID** | H2020-SC6-MIGRATION-2018-2019-2020 – DT-MIGRATION-06-2018 |
| **Work Package** | *WP3 – Data Analysis and skills matching* |
| **Work Package Leader** | *Universidad Politécnica De Madrid UPM* |
| **Deliverable Leader** | *Universidad Politécnica De Madrid UPM* |
| **Deliverable coordinator** | Gustavo Hernández ghp@gatv.ssr.upm.es |
| **Deliverable Nature** | Prototype |
| **Dissemination level** | Public (PU) |
| **Version** | 1.3 |
| **Revision** | Final |

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

# Document Info

## Authors


| Author name | Organization | E-Mail |
|---|---|---|
| David Martín Gutierrez | UPM | dmz@gatv.ssr.upm.es |
| Gustavo Hernández | UPM | ghp@gatv.ssr.upm.es |


## Document History

| Version # | Author name | Date | Changes |
|---|---|---|---|
| 0.1 | David Martín | 18-02-2020 | Starting version |
| 0.2 | Gustavo Hernández | 1-03-2020 | Initial Contributions |
| 0.3 | David Martín | 09-03-2020 | Contributions in section 3 |
| 0.4 | David Martín | 12-03-2020 | Extension of section 2 |
| 0.5 | David Martín | 29-03-2020 | Extension of section 3 |
| 0.6 | Gustavo Hernández | 03-04-2020 | Complete version Review |
| 0.7 | Thodoris semertzidis | 13-04-2020 | Peer review |
| 0.8 | Pilar Orero | 16-04-2020 | Internal peer review |
| 0.9 | Annelore Hermann | 20-04-2020 | External revision |
| 1.0 | Antonio Salvador Filograna | 21-04-2020 | Complete version review |
| 1.1 | Vukasin Simic | 22-04-2020 | Complete External review |
| 1.2 | David Martín | 23-04-2020 | Final remarks and complete review |
| 1.3 | Gustavo Hernández | 23-04-2020 | Final submitted version |

## Document Data

| | |
|---|---|
| **Keywords** | D3.2 Methodology for skills and needs matching |
| **Editor Address data** | Name: David Martín, Gustavo Hernández<br>Partner: UPM<br>Address: Av Ramiro de Maeztu 7<br>Phone: +34 91 464160 ext. 142<br>Email: [dmz,ghp]@gatv.ssr.upm.es |
| **Delivery Date** | 15-04-2020 |
| **Peer Review** | *thodoris semertzidis CERTH* |

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

# EXECUTIVE SUMMARY

This deliverable is reporting the work performed and followed in REBUILD project's WP3 to provide a skill-matching methodology based on artificial intelligence techniques. More specifically, this procedure relies on the user's profile analysis computed in Task 3.1 to associate specific needs of each user with a certain service or tool available in the REBUILD platform.

The scope of tasks in WP3 is totally interrelated, where the profiles embedding done in Task T3.1 (and reported in D3.1) is inserted in the module presented here, to provide end-2-end matching (with an end ranking list) based on the inputs. This information will be employed also by the recommendation engine (T3.3) to provide its recommendations.

Moreover, the scope of this deliverable is the description of the methodologies and procedures that are performed within the REBUILD project to match the needs of end-users or migrants with the services and tools that are available in the REBUILD platform according to their profiles. Therefore the input of this component is the one provided in Task 3.1 where the user profile is embedded into a more compact representation that is passed throughout the skill-matching functionality in order to retrieve the more adequate services to the end-user.

Since REBUILD is a user-driven project the information of this deliverable will be updated with the decisions taken within the co-creation and the other REBUILD workshops with migrants and stakeholders to better approach their actual needs.

For this analysis, the European Skills, Competences, Qualifications and Occupations Ontology (ESCO) as reference for the jobs classification and how this ontology reflects the needs for this particular topic was used.

The document is organised as follows: The first section is devoted to review the literature and select the most appropriate approaches for skill matching. It involves the analysis of the most relevant dataset structure and the techniques (supervised and unsupervised) for the overall skills matching topic. This analysis allows us to understand what is the processing pipeline and also to use some existing ontologies (adapted to the REBUILD needs). Then, the second section is intended to describe the most relevant techniques, as well as the implementation details. The technologies employed for the development of this module are provided, as well as the details of the interaction and goals.

Finally, technical details of the physical deployment of this module, involving libraries and dockerization are provided. This deliverable has been submitted with a minor delay due to the unexpected situation derived from the COVID-19.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

# Table of Contents

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

# INDEX OF FIGURES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

# 1 INTRODUCTION

REBUILD project's objective is to provide a toolbox of ICT-based solutions that will help to the smooth integration of refugees and migrants. REBUILD refers both to the facilitation of the local authorities' management procedures and to the migrants' life quality improvement. To achieve these goals, REBUILD is designed as a user-centered application that attempts to recognize users' needs and give them personalized recommendations and targeted solutions. To assess this purpose, personal information for each migrant is required in order to learn profile patterns and link to needs and resources. For the gathering of those necessary data, all users will have to consent in order to provide anonymized, GDPR-compliant information that will be used by AI-based methods. In particular, the proposed technological solutions include an AI-based profile analysis to enable the personalized support, an AI-based matching tool in order the migrants' needs and skills to be matched with services provided by local authorities in each pilot country and a set of tools such as a chatbot or audio visual communication to enable personalized two-way, effective communication between the final users, i.e migrants and local service providers.

More specifically, this project follows a user-centered and participated design approach, aiming at addressing properly real target users' needs, ethical and cross-cultural dimensions, and at monitoring and validating the socio-economic impact of the proposed solution. Both target groups (immigrants/refugees and local public services providers) will be part of a continuous design process; users and stakeholders' engagement is a key success factor addressed both in the Consortium composition and in its capacity to engage relevant stakeholders external to the project. Users will be engaged since the beginning of the project through interviews and focus groups; then will be part of the application design, participating in three Co-Creation Workshops organized in the three main piloting countries: Italy, Spain and Greece, chosen for their being the "access gates" to Europe for main immigration routes. Then again, in the 2nd and 3rd years of the project, users' engagement in Test and Piloting events in the three target countries, will help the Consortium fine-tuning the REBUILD ICT toolbox before the end of the project.

The key points regarding technology solutions proposed are:

- GDPR-compliant migrants' integration related background information gathering with user consent and anonymization of personal information;
- AI-based profile analysis to enable both personalized support and policy making on migration-related issues;
- AI-based needs matching tool, to match migrant needs and skills with services provided by local authorities in EU countries and labour market needs at local and regional level;
- a Digital Companion for migrants enabling personalized two-way communication using chatbots to provide them smart support for easy access to local services (training, health, employment, welfare, etc.) and assessment of the level of integration and understanding of the new society, while providing to local authorities data-driven, easy to use decision supporting tools for enhancing capacities and effectiveness in service provision.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

As it was mentioned above, this deliverable is focussed on the AI-based tools that are provided within the project in order to improve the Quality of Experience (QoE) of users.

# 2 Skill Matching: Background and State Of the Art

## 2.1 Introduction

The Skill-matching framework has gained more attention in recent years since it has become an essential tool in many applications such as job seeking, which is widely used in both Human Resources management and online recruitment.

Thus, essentially, the goal of this framework is to facilitate web recruitment applications by searching and finding automatically the best job-seekers for a given job as well as positions for a certain job-seeker.

Moreover, a skill-matching component usually has three main parts: an input, a latent information and the target. As an example, considering the job seeking application, the inputs are candidates, the latent information are their skills and the target are jobs. As expected, there is a hidden relationship between inputs and targets which are indeed the skills.

However this problem can be easily extrapolated to additional applications such as house seeking (accommodation), e-learning among many others. Additionally, this module is normally used as a stage within a recommender system which can have different outputs such as job seeking and accommodation, event recommendations depending on the final application.

## 2.2 State Of The Art in Skill-Matching

Skill-matching is a wide branch of research in recent years where multiple approaches are being followed. More specifically, there exist mainly four families of methodologies including logic description, machine learning, semantic ontology-based approaches as well as hybrid procedures as it is suggested in (Petrican, T., Stan et al. 2017).

More specifically, authors in (Calì, A., Calvanese et al, 2004) suggest a Description Logic (DL) system which uses logical frameworks and algorithms to attempt to match demands and needs of profiles considering incompleteness of user profiles. The final solution of this approach is employed in both recruitment and dating applications.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

Another perspective when implementing a skill-matching engine is adding machine learning procedures as authors suggest in (Botov, D., Klenin, et al. 2019) to build a system capable of automatically extracting a collection of objective criteria from user's profiles and then infer their personality competencies by the use of potential linguistic features. However, these approaches require a considerable amount of data in order to properly train models that assess the problem with a high level of precision.

Moreover, in (Sayfullina, L., Malmi, et al. 2018), authors propose a phrase-matching-based approach which attempts to differentiate between what they denoted as soft skills phrases referring to a given candidate vs something else. This disambiguation is then determined as a binary text classification problem where the goal is to predict the potential soft skills based on the context where it occurs. Additionally authors suggest two main approaches including soft skill *masking* and soft skill *tagging*.

On the other hand, Semantic ontology based approaches are the most widely used in the state-of-art and they have been analysed in several investigations as the ones proposed in (Hohaghegh, S., & Razzazi, M. R., 2004), (Bizer, C., Heese, R., Mochol 2005), (Hassan, F. M., Ghani, I. et al., 2012) and (Petrican, T., Stan et al. 2017). These four research works address the problem by modelling users, competencies and jobs using an ontology and subsequently, applying specific graph algorithms to compute rankings and similarities between jobs and users based on the competencies that they may have in common.

Finally, many investigations are following a Hybrid solution where the idea is to combine either description logic or machine learning approaches on top of a semantic ontology methodology. In particular, authors in (Fazel-Zarandi, M. et al. 2009) present a DL solution on top of an ontology approach where the goal is to use the DL system together with rules in order to represent and reason about both applications and job advertisements. On the other hand, authors in (Doan, A., Madhavan, et al. 2004) develop a system named as GLUE which uses learning techniques to build a collection of semantic mappings between different ontologies via semi-automatically procedures.

# 3 Skill Matching Implementation Methodology

The scope of this section consists in describing the methodology that is followed to assess the skill-matching procedure within the REBUILD project. As it was already mentioned in previous sections the goal of a skill-matching system is to associate user profiles to specific demandings or supplies that may facilitate the professional life of both end-users and stakeholders by automatically performing this functionality. More specifically, within the REBUILD scenario, the skill-matching system is mainly focused on job seeking. However, it will be used as an input of the recommendation engine which is in charge of providing specific services such as HealthCare, Housing or Social events based on the user profiles as well.

Moreover, several families of approaches were presented in the previous section in order to introduce the distinct ways that can be followed to achieve the desired purpose. However, the Hybrid solution is the one preferred to be used in this scenario since it relies on both a semantic ontology to represent

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

the main entities of the problem, and machine learning techniques to perform the ranking of the more adequate jobs or services in an automatic way.

It is important to remark that skill-matching algorithms and techniques are widely used within the job seeking framework where we have some target information (jobs), a given information (candidates) and a latent information (skills). Therefore, the goal in this case consists in associate jobs to candidates or considering a reverse procedure, associate candidates to jobs.

However, it is possible to extend the skill-matching procedure through different scenarios such as social education or housing which are services that may be incorporated in the whole REBUILD platform.

## 3.1 Skill-Matching Ontology

### 3.1.1 European Skills, Competences Qualifications and Occupations Ontology

The first step of the process consists in building the ontology of the application. To do so, we have firstly analysed the so-called ESCO[1] (European Skills, Competences, Qualifications and Occupations) which is a European Commission project and it has an Application Programme Interface (API) to be used by open services and applications. In particular, ESCO contains a collection of 2942 occupations and 13.485 skills which are linked to these occupations. Additionally, it is translated into 27 languages (all official EU languages plus Icelandic, Norwegian and Arabic).

Moreover, the dataset of the ESCO can be downloaded and customized according to the final application using this link[2]. It allows developers to select the data they desire in specific languages according to their needs.

Furthermore, in order to map the data provided by ESCO in separated files into a graph representation, we firstly need to establish the entities that are involved as well as their relationships between each other. More specifically, the following entities are mapped regarding the ESCO dataset:

- **Occupation**: this object will represent the different occupations provided in the ESCO. Each of them has a unique identifier that is used to avoid duplications during the process of ingesting data into the graph.
- **ISCOGroup**: this object will indicate the category of the Occupation object. It also contains a unique identifier.
- **Skill**: this entity or object will represent the set of skills provided in the ESCO dataset. As in previous cases, each skill will contain a specific identifier to be unique in the whole representation.

[1] https://ec.europa.eu/esco/portal
[2] https://ec.europa.eu/esco/portal/download

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

- **SkillGroup**: this entity will indicate the parent group of the different Skill objects in order to establish a certain hierarchy within the skills representation.

Moreover, each of these entities contains different properties which are presented in the following tables:

| Property | Type | Example |
|---|---|---|
| iscoGroup | integer | 2166 |
| conceptUri | url | http://data.europa.eu/esco/occupation/00030d09-2b3a-4efd-87cc-c4ea39d27c34 |
| inScheme | url | http://data.europa.eu/esco/concept-scheme/occupations, http://data.europa.eu/esco/concept-scheme/member-occupations |
| preferredLabel | string | technical director |
| conceptType | string | Occupation |
| modifiedDate | UTC Date | 2016-07-05T13:58:41Z |
| description | string | Technical directors realise the artistic visions of the creators within technical constraints (...) |
| altLabels | string | technical and operations director<br>head of technical |
| regulatedProfessionNote | url | http://data.europa.eu/esco/regulated-professions/unregulated |
| status | string | released |

*Table 1: Data model for the Occupation entity of the ESCO dataset.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

| Property | Type | Example |
|---|---|---|
| code | integer | 0 |
| conceptUri | url | http://data.europa.eu/esco/isco/C0 |
| inScheme | url | http://data.europa.eu/esco/concept-scheme/occupations, http://data.europa.eu/esco/concept-scheme/isco |
| preferredLabel | string | Armed forces occupations |
| conceptType | string | ISCOGroup |
| description | string | Armed forces occupations include all jobs held by members of the armed forces. (...) |

*Table 2: Data model for the ISCOGroup entity of the ESCO dataset.*

| Property | Type | Example |
|---|---|---|
| conceptUri | url | http://data.europa.eu/esco/skill/25a26ff6-af18-40b0-b2e2-cb58471015eb |
| preferredLabel | string | values |
| conceptType | string | SkillGroup |
| description | string | Principles or standards of behaviour, revealing one's judgement of what is important in life. (...) |
| altLabels | string | scruples<br>beliefs |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

| | | |
|---|---|---|
| | | morals" |

*Table 3: Data model for the Skillgroup entity of the ESCO dataset.*

| Property | Type | Example |
|---|---|---|
| skillType | string | skill/competence |
| reuseLevel | string | sector-specific |
| conceptUri | url | http://data.europa.eu/esco/skill/0005c151-5b5a-4a66-8aac-60e734beb1ab |
| inScheme | List of urls | http://data.europa.eu/esco/concept-scheme/skills, http://data.europa.eu/esco/concept-scheme/member-skills |
| preferredLabel | string | manage musical staff |
| conceptType | string | KnowledgeSkillCompetence |
| modifiedDate | UTC Date | 2016-07-05T13:58:41Z |
| description | string | Assign and manage staff tasks in areas such as scoring, arranging, copying music and vocal coaching. (...) |
| altLabels | string | manage staff of music<br>coordinate duties of musical staff<br>manage music staff |
| regulatedProfessionNote | url | http://data.europa.eu/esco/regulated-professions/unregulated |
| status | string | released |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

*Table 4: Data model for the Skill entity of the ESCO dataset.*

Moreover, in order to visualize the ontology as a graph representation, we have employed NEO4J[3] which is a powerful graph framework with the functionality of mapping different ontology representations such as rbf, csv or any other file into a graph representation using its own query language named Cypher.

Consequently, we used Cypher to proceed with the mapping among the ontology files regarding the ESCO dataset in order to build the graph as it is suggested in this Neo4j blog[4]. In particular, the following instructions are needed:

---

```
create index ON :Occupation(ISCOGroup);
create index ON :Occupation(altLabels);
create index ON :Skill(altLabels);
create index ON :ISCOGroup(code);
create index ON :Skill(conceptUri);
create index ON :ISCOGroup(conceptUri);
create index ON :Occupation(conceptUri);
create index ON :Occupation(preferredLabel);
create index ON :Skill(preferredLabel);

//import skills and skillgroups
//skillgroups are also skills
load csv with headers from "file:///skillGroups_en.csv" as row
create (s:Skill:Skillgroup)
set s = row;
//skills
load csv with headers from "file:///skills_en.csv" as row
create (s:Skill)
set s = row;
//add the BROADER_THAN relationship between different skills
load csv with headers from "file:///broaderRelationsSkillPillar.csv" as row
match (smaller:Skill {conceptUri: row.conceptUri}), (broader:Skill {conceptUri: row.broaderUri})
create (broader)-[:BROADER_THAN]->(smaller);
```

[3] https://neo4j.com/
[4] http://blog.bruggen.com/2018/08/esco-database-in-neo4j-skills.html

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

```
//import occupations
load csv with headers from "file:///occupations_en.csv" as row
create (o:Occupation)
set o = row;

//import the International Standard Classification for Occupations of the ILO
load csv with headers from "file:///ISCOGroups_en.csv" as row
create (isco:ISCOGroup)
set isco = row;
//import the BROADER_THAN relationships between ISCO groups
load csv with headers from "file:///broaderRelationsOccPillar.csv" as row
match (smaller:ISCOGroup {conceptUri: row.conceptUri}), (broader:ISCOGroup {conceptUri: row.broaderUri})
create (broader)-[:BROADER_THAN]->(smaller);
//connect the occupations to their ISCOGroup
match (isco:ISCOGroup), (o:Occupation)
where isco.code = o.iscoGroup
create (o)-[:PART_OF_ISCOGROUP]->(isco);

//Connect Skills to Occupations
using periodic commit 500
load csv with headers from "file:///occupationSkillRelations.csv" as row
match (s:Skill {conceptUri: row.skillUri}), (o:Occupation {conceptUri: row.occupationUri})
CREATE (s)-[:RELATED_TO {type: row.relationType}]->(o);

// match ()-[r:RELATED_TO]->()
// return distinct r.type

//differentiate the different types of relations between occupations and skills
match (a)-[r:RELATED_TO]->(b)
where r.type = "essential"
create (a)-[:ESSENTIAL_FOR]->(b);
match (a)-[r:RELATED_TO]->(b)
where r.type = "optional"
create (a)-[:OPTIONAL_FOR]->(b);
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

```
//remove the old relationships
match (a)-[r:RELATED_TO]->(b)
delete r;
```

---

An example of this kind of visualization is depicted in Figure 1, where a part of the graph regarding the ESCO ontology is presented. In particular, by observing the graph one may check a set of skills (in red) that are associated with certain occupations (in blue) which are also associated with specific ISCOgroups (in green) as described in table 4.

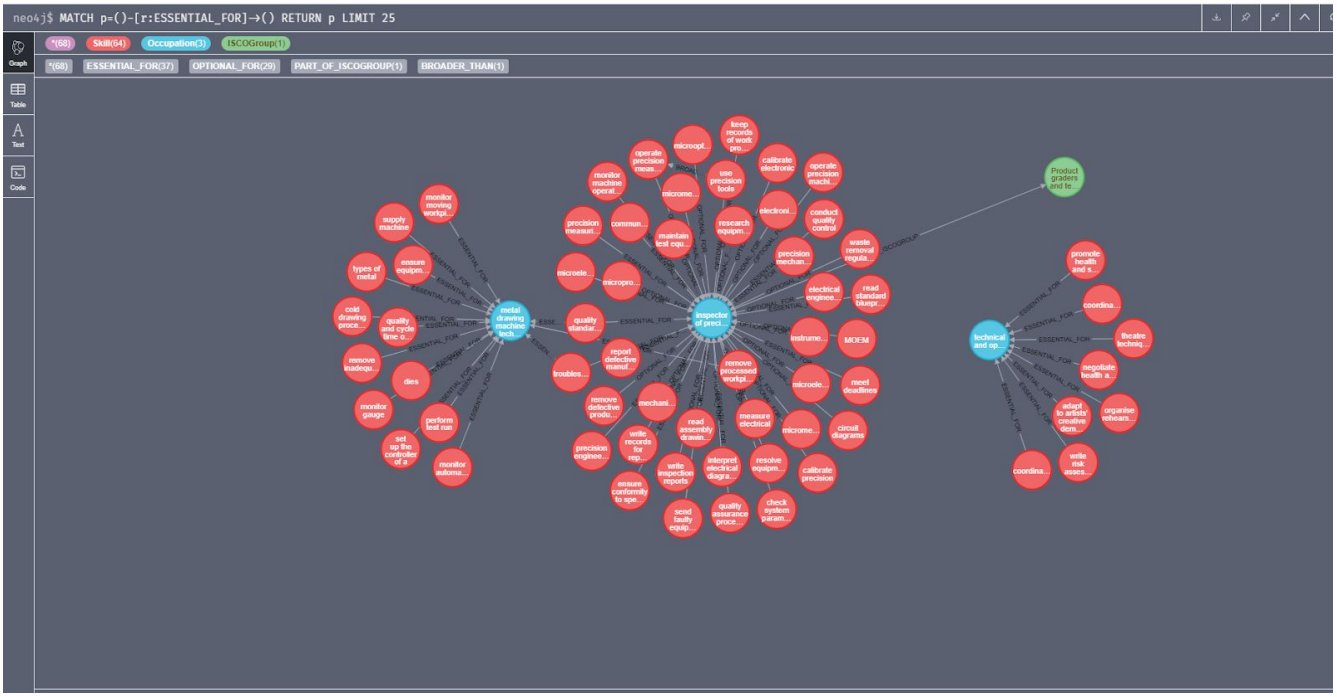


*Figure 1: A sub-graph visualization representing different entities of the ESCO ontology including Skills (red), occupations (blue) and ISCOgroup (green).*

Finally, using the NEO4J graph representation, we are able to perform different algorithms including similarity and clustering approaches, centrality and importance rank approaches as well as short-path methods to facilitate the matching between occupations and skills.

### 3.1.2 Simplified REBUILD Ontology For Skill Matching

The aforementioned ontology constitutes a complete representation of the European perspective regarding relationships between pairs of occupations and skills. However, an adaptation of this complete ontology must be addressed in order to be included within the REBUILD platform, where a simplification of both skills and occupations is required. The aim is to both facilitate the integration of

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

migrants in arrival countries as well their interaction with the platform and to bound the amount of information demanded from them.

Moreover, to establish the REBUILD ontology, a minimum quantity of data is needed in order to have both the entities that will be related to each other as well as the set of properties that each entity will have. Therefore, it is important to remark that many of the properties and entities that are presented in the next lines would not be the final ones since some of the suggested information may not be available or may not be allowed to be asked in order to comply with the GDPR terms. In addition, there will be an updated version of this document at the end of the project which will have the final version of the needed skill-matching framework used within REBUILD.

Therefore, bearing in mind the aforementioned considerations, the following entities are included in the analysis:

- **User**: the final user of the REBUILD application, which can be, migrants, service providers or any other stakeholder involved in the process.
- **Skill**: this entity represents the competencies that a user needs to perform a certain job.
- **Job**: this entity denotes a job object that will be uploaded by job recruiters to be used by the skill matching to infer the best jobs associated with each User of type migrant.
- **Topic**: this entity will represent interests, needs and/or hobbies that a certain User may have.
- **Content:** this entity represents the material, content or activities that stakeholders and service providers will upload to the platform to be recommended to the final user.

This list of entities will be increased when all the services of the platform are ready to be deployed. Then, the rest of the elements that will be needed to complete the ontology of the project will be included.

Furthermore, a set of relationships among the different entities is also required to build the final graph. Hence, we need to define the following relationships to be incorporated to the NEO4J graph:

- **IS_INTERESTED_AT**: represents a relationship between entities User and Topic.
- **IS_NEEDED_FOR**: represents a relationship between entities Skill and Job.
- **HAS_EXPERTISE_IN**: represents a relationship between entities User and Skill.
- **IS_TAGGED_AS**: represents a relationship between entities Content and Topic.

Having all these considerations, a graph can be built in order to have all the entities related according to the entity type and their corresponding relationships.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

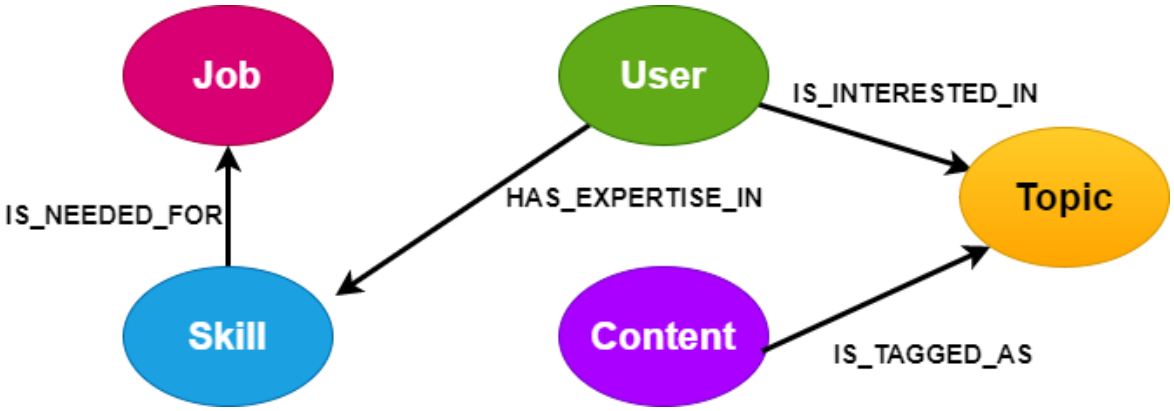


*Figure 2: Visual representation of the different entities that are involved in the REBUILD ontology for skill-matching purposes.*

In Figure 2 plots in a visual representation the aforementioned entities that are included in the current ontology for skill-matching purposes. Furthermore, the collection of properties associated with each entity is also needed, so that the following tables contain the properties as well as their descriptions for each of the involved components of the ontology. Moreover, the asterisk (*) indicates that the string will be encrypted using hashing procedures in order to comply with the GDPR terms.

| Property | Type | Example |
|---|---|---|
| username | string | r09garcia |
| email | string | r09garcia@hotmail.com |
| age (*) | string | 18-30 |
| Mother_tongue (*) | string | english |
| Studies_level (*) | string | primary |
| Origin_country (*) | string | Morocco |
| Arrival_country (*) | string | Spain |
| Interests (*) | Lists of strings | ["music", "culture", "sports", "night"] |
| uuid | string | 51a10467639511bf8e49a03eaf5d671135670b4bfd82be58e40922c6f04ee53455fb1e1c5abb597ec2d9232e74b6db9ce2cf2fdcdda22903b0ae15cd9b439b77 |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

| | | |
|---|---|---|
| skills_uuids | List of strings | [bfbe0487366baee969d9c6e9c1b587a14701520a19fc9bb0bb7768942e0232f07a9da11449fd37f23422e3eb26d91d8ebd9821692ba71521841949 7b88599f29, 0a8be2ca26e6e32c40f4e454fcda683b9f0031d1ead4ad1c6fba4715fad028032c94856c6b4ef57d4e3e0f8d68cfd8f4e83adab7f032ab82ba36408153b37941] |
| skills_level | List of integers | [4,2] |

*Table 5: Data model for the User entity of the REBUILD ontology for the skill-matching component. The asterisk (*) indicates that the values will be encrypted by hashing them in order to comply with the GDPR terms.*

| Property | Type | Example |
|---|---|---|
| name | string | communications and media |
| description | string | knowledge of media production, communication, and dissemination techniques and methods. this includes alternatives (...) |
| skill_type (*) | string | knowledge |
| normalized_skill_name | string | communications_and_media |
| uuid | string | 703862951cf4bbf9e4e58954127df668f5545fb4ec60ee9dee1f6e99fc00f1d5378a747e354723335ef73014c50999ad2df2943f815a5820d0080de403a6cc6e |

*Table 6: Data model for the Skill entity of the REBUILD ontology for the skill-matching component. The asterisk(*) indicates that the values will be encrypted by hashing them in order to comply with the GDPR terms.*

| Property | Type | Example |
|---|---|---|
| title | string | 3D Modeler |
| normalized_job_title | string | 3d_modeler |
| description | string | knowledge |
| skills_uuids | list of strings | [703862951cf4bbf9e4e58954127df |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

| | | |
|---|---|---|
| | | 668f5545fb4ec60ee9dee1f6e99fc00f1d5378a747e354723335ef73014c50999ad2df2943f815a5820d0080de403a6cc6e, 9a8b24f8b81c6a921ef5d3e16ad83730e96ac766d3239e4d76faf677a9faf804816376f43f9b0d97142e2cf62d6bafbe08e12ce3de37b7c882f2fae10c186f5d] |
| skills_importances | list of integers | [4,4] |
| uuid | string | 703862951cf4bbf9e4e58954127df668f5545fb4ec60ee9dee1f6e99fc00f1d5378a747e354723335ef73014c50999ad2df2943f815a5820d0080de403a6cc6e |
| localization | list of floats | [37.4224764, -122.0842499] |

*Table 7: Data model for the Job entity of the REBUILD ontology for the skill-matching component. The asterisk(*) indicates that the values will be encrypted by hashing them in order to comply with the GDPR terms.*

| Property | Type | Example |
|---|---|---|
| name | string | Go to concerts |
| normalize_name | string | go_to_concerts |
| type | string | cultural |
| uuid | string | 103822951cf4bbf9e4e58954127df668f5545fb4ec60ee9dee1f6e99fc00f1d5378a747e354723335ef73014c50999ad2df2943f815a5820d0080de403a6cc6p |

*Table 8: Data model for the Topic entity of the REBUILD ontology for the skill-matching component. The asterisk (*) indicates that the values will be encrypted by hashing them in order to comply with the GDPR terms.*

| Property | Type | Example |
|---|---|---|
| name | string | Things to do in Madrid |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

| | | |
|---|---|---|
| description | string | Let's explore the best things to do in Madrid for this weeked! |
| tags | list of strings | [cultural, music, events] |
| uuid | string | 103822951cf4bbf9e4e58954127df668f5545fb4ec60ee9dee1f6e99fc00f1d5378a747e354723335ef73014c50999ad2df2943f815a5820d0080de403a6cc6p |
| link | url | https://eventful.com/madrid/events |

Table 9: Data model for the Content entity of the REBUILD ontology for the skill-matching component. The asterisk (*) indicates that the values will be encrypted by hashing them in order to comply with the GDPR terms.

The aforementioned information provided in the tables will be updated according to the general REBUILD data model of the platform as well as the concerns of the GDPR terms in case some of the above properties may be sensitive to be stored.

Furthermore, in order to run some initial experiments, a data generation process was needed to generate synthetic users. On the other hand, by using external skill-jobs APIs such as the one provided by "data works" [5], a considerable number of jobs and skills were gathered.

## 3.2 Skill-Matching Dataflow

The scope of this section consists in describing the general data flow that the skill-matching component requires to incorporate its functionality into the general REBUILD application. Moreover, a use case regarding job seeking will be analyzed in order to incorporate more details of the process in a more particular scenario or use case.

1. Register Process
   - 1.1. A migrant is registered in the application by filling out a brief form where data is asked (Table 5: Data model for the User entity of the REBUILD).
   - 1.2. A third party is registered in the application by filling out a brief form where data is asked (Table 5: Data model for the User entity of the REBUILD).
2. User Data storage process
   - 2.1. All the data from step 1) is directly stored in the general database of the platform for further analysis.
   - 2.2. All the data from step 1) is directly stored in the skill-matching database of the platform for further analysis.

[5] http://api.dataatwork.org/v1/spec/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

3. Put new content in the platform (i.e jobs)
   3.1. A third party logins in the system and incorporates information about new content (i.e jobs) using a form.
4. Content Data storage process
   4.1. All the data from step 3) is directly stored in the general database of the platform for further analysis.
   4.2. All the data from step 3) is directly stored in the skill-matching database of the platform for further analysis.
5. Skill-matching request
   5.1. A component triggers a request to the Skill-matching component to extract some knowledge from a given user uuid (i.e the best K jobs more adequate for this user uuid).
6. Skill-matching process
   6.1. The skill-matching engine starts computing the techniques and the algorithms to retrieve the information that was requested.
7. Skill-matching response
   7.1. The skill-matching engine returns a collection of sorted uuids (i.e best jobs for the user) to the component that sent the request.

In the following figure (Figure 3), a visual block diagram shows the aforementioned data flow for the skill-matching engine or component in a particular scenario: job seeking, which is one of the most important and relevant services within the REBUILD platform which will facilitate migrants to be incorporated in the laboral market according to their skills and competencies both personal and professional.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

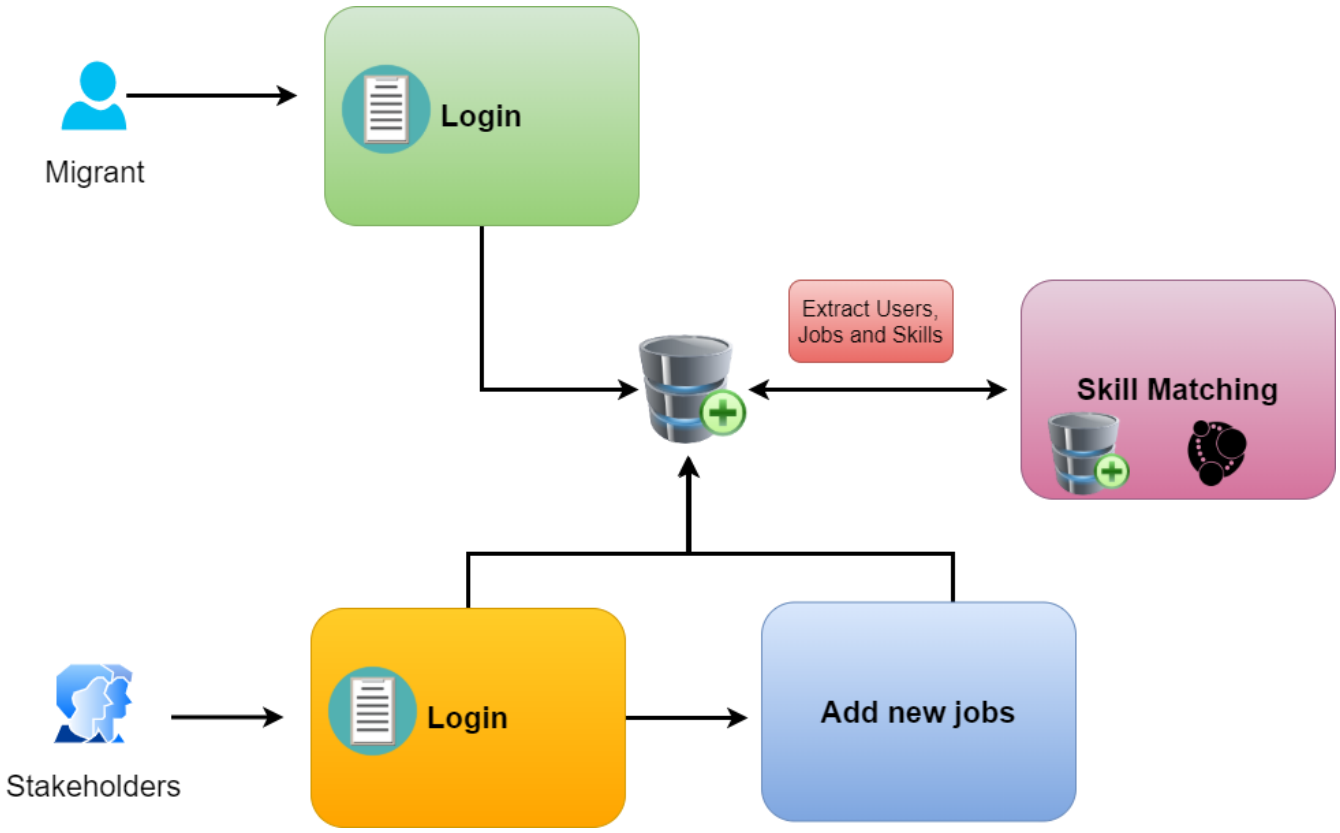


*Figure 3: Block diagram schema representing the workflow of the job-seeking service using the skill matching service.*

As one may observe in Figure 3, the skill-matching component contains its own database using NE4J graph framework, where all the entities needed for the analysis will be related using the criteria that were described in the previous section.

Moreover, the main objective of having this local database lies in avoiding the skill-matching process to access the whole data of the general database every time it needs to process the information of a certain user. This neo4j db will speed the process, and will not be a problem in terms of data integrity, as it will contain a semi processed information that will be synchronised with the general ReBuild DB. In this way, every time that new information is incorporated to the system as it is needed to the skill-matching component, it will be also copied into this local database.

Furthermore, in Figure 4, a visual representation of the graph database stored inside the Skill-matching component is shown. As it is observed, three elements of the simplified REBUILD Ontology are presented using the aforementioned relationships including: User, Skill and Job entities.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

In the previous sections, we mentioned the fact that having a graph structure is more adequate to perform any other semantic analysis afterwards. More specifically, some procedures such as user profiling (clustering), job/skill/user importance (centrality algorithms) or even more advanced machine learning techniques can be computed at the top of this architecture.

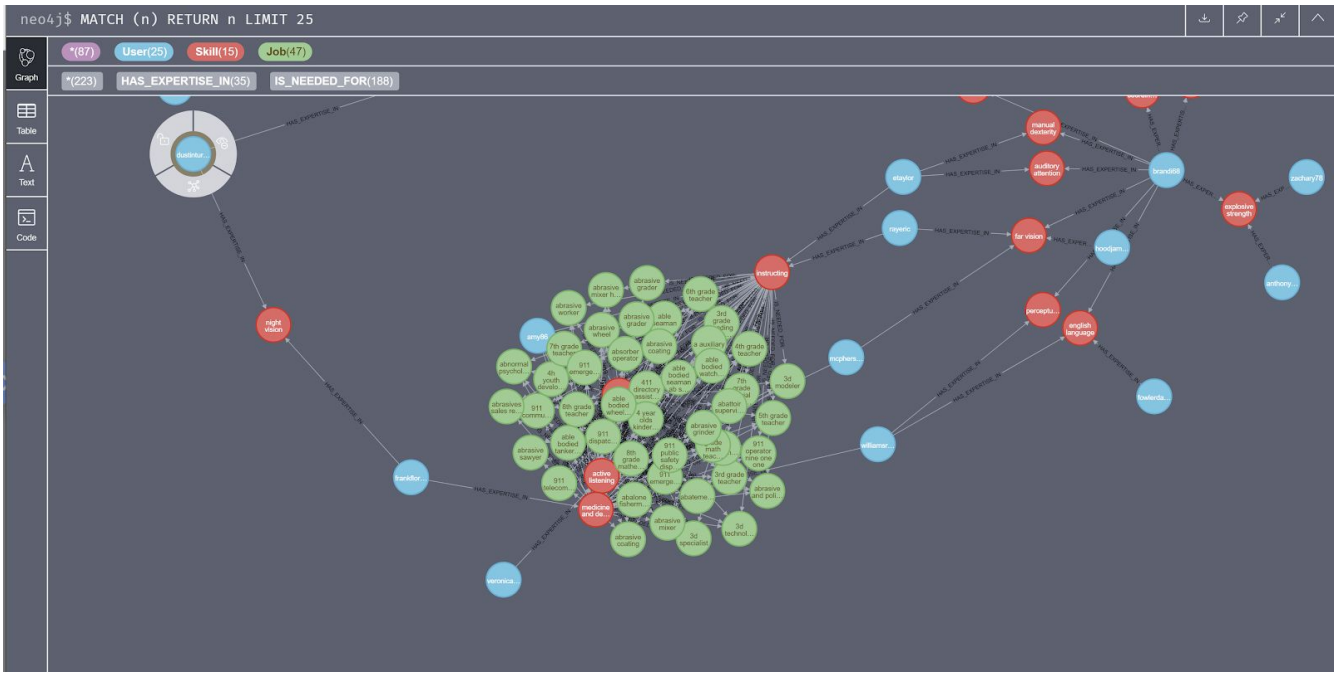


*Figure 4: Graph visualization showing an example of a set of Users (in blue), Skills (in red) and Jobs (in green) including the relationships between each other.*

## 3.3 Skill-Matching Techniques For Job Seeking

Along this section, a detailed review of the main techniques that are used, or considered to be used and improve in future releases, regarding the skill-matching component is presented. In particular, we will focus the analysis on the Job Seeking scenario since it is the most common when talking about this procedure.

Firstly, it is necessary to remark that generally, the job seeking flow can be described by answering the following question:

- Given a job, what is the probability of a certain candidate to be adequate enough for such a job?

The following figure illustrates the main idea where variable *x* represents the job, variable *y* represents the candidate whereas *z* indicates the latent factors.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

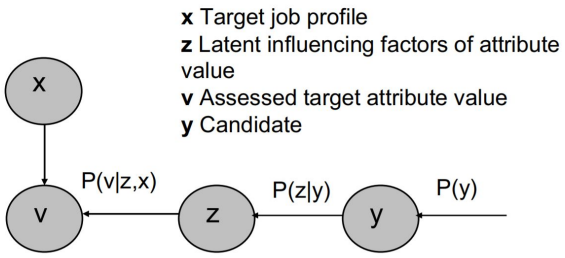


*Figure 5: Visual representation of the probabilities involved in the process of job recommendation. Extracted from Malinowski, J., Keim, et al. (2006).*

However, the proposed scenario in REBUILD is exactly the opposite version:

- Given a candidate or user, what is the probability that a certain job is adequate for him/her?

Therefore, the idea here is to associate, from a large list of jobs available, which are the most adequate ones for each available user in the platform at a particular period of time.

The first technique to perform a skill-matching is based on what authors in (Malinowski, J., Keim, et al. 2006) denote as the person-job fit conceptualization, which basically consists in measuring how adequate a certain person or user is for a particular job based on their competences. To do this, the simplest solution lies in analysing the intersection between the user skills and the job skills and compute a ranking to measure indeed how adequate the user is. In this case, all the skills have the same weight or importance when applying for a job.

An adaptive version of the aforementioned technique consists in adding different weights to each skill in the sense that, users may incorporate a certain level of knowledge or expertise on each field or skill and besides, candidate recruiters may also include these importances when uploading a new job within the platform. Thus, the procedure of measuring the level of "fitness" of the jobs for a given candidate has now a more precise outcome by incorporating the importances associated with the set of skills.

Thus, these solutions follow an Ontology-based approach as the one described at (Petrican, T., Stan, et al. , 2017), where a given user $\Omega_i \forall i=1,..,i=M$ is represented as a node in a graph of skills and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

all the relationships represent the set of skills associated to $\Omega_i$. We denote such a set of skills as $S_\Omega = \{s_1, s_2, \ldots, s_K\}$.

Moreover, a set of weights $W_{(\Omega,S_\Omega)} = \{w_1, w_2, \ldots, w_K\}$ is employed to measure the importance of each skill. In this case, this importance value indicates the expertise that a given user $\Omega_i$ has for each $s_k$.

Additionally, a set of jobs $\Theta = \{\theta_1, \theta_2, \ldots, \theta_P\}$ is also involved in the graph representation, where each particular $\theta_p$ represents a node which contains a set of required skills $S_\Theta$ with their corresponding set of weights $W_{(\Theta,S_\Theta)}$ associated with them as well. All the weights are normalized to sum up to 1 in order to provide a final matching score between 0 and 1.

Hence, the algorithm attempts to search and match the subset of jobs that shares the highest number of skills with the given user considering also the weights to measure the level or degree of matching.

A visual representation of this procedure is shown in *Figure 6,* where Users, Skills and Jobs are depicted as nodes in what is normally called Skill-based Graph. Where users are the input, the skills are the latent information and finally, jobs are the target of the system.

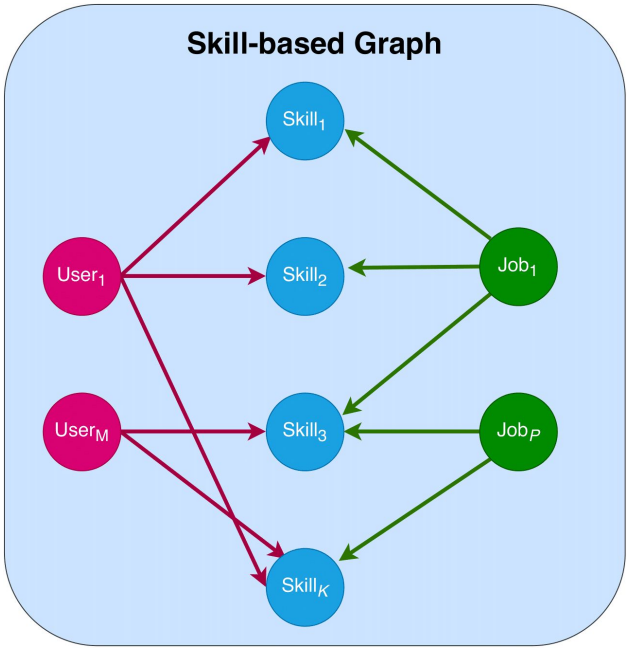


*Figure 6: Visual representation of the skill-based graph where the set of users $\Omega$, skills $S$ and jobs $\Theta$ are presented as red, blue and green circles respectively. The set of weights $W_{(\Omega,S_\Omega)}$ is shown as red lines whereas the set of weights $W_{(\Theta,S_\Theta)}$ is represented with green lines.*

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

The aforementioned techniques are user-item approaches where the final solution has not considered the rest of the users in order to provide the most adequate jobs for a particular user. However, there are other approaches that take advance of the profiles of similar users in order to provide the final jobs.

More specifically, the idea is the following: by analyzing a first set of users, a list of jobs will be associated with them using the previous techniques. However, when the set of users is big enough, it is more appropriate to perform a clustering or profiling and then provide or recommend the same jobs to those users that belong to the same class or cluster. By doing this, we avoid performing the recommendation or matching for one user each time and therefore, given a user, the system will compute the best jobs not only for this user but for those who belong to the same category.

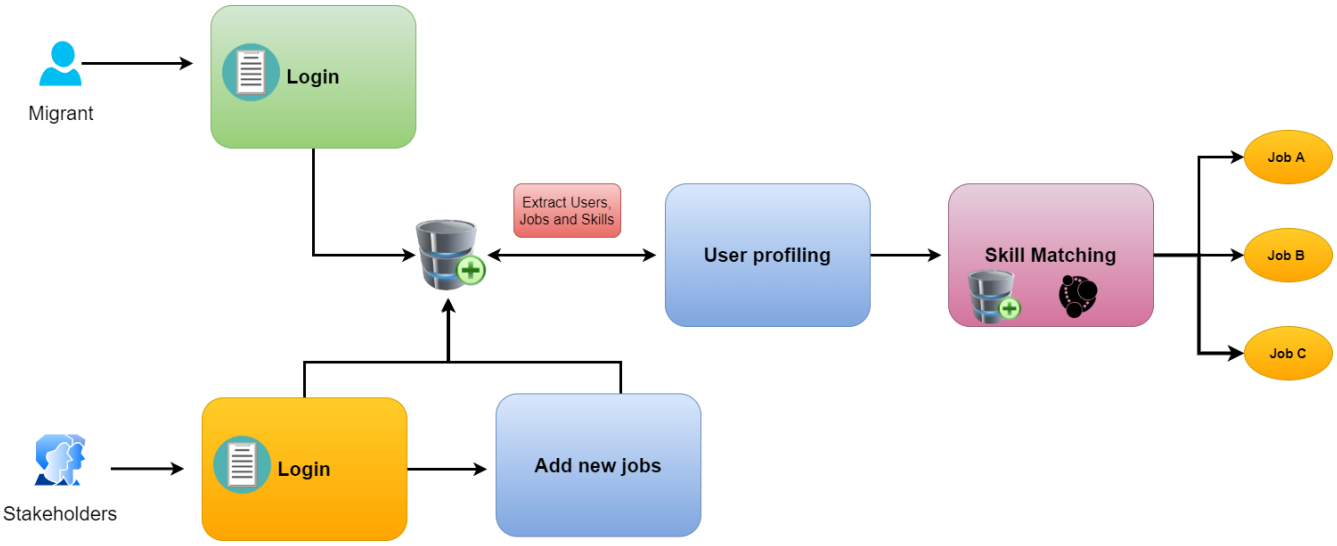


Figure 7: Block diagram showing the skill-matching process when using user profiling as input of the system.

In the figure above, one may observe that now the input of the skill-matching is the user profiling output, which basically will be a set of embeddings as well as their associated classes or clusters. Therefore, when a new migrant is registered in the application and the skill-matching is launched, a new embedding is computed for this migrant using the trained model and it will have a cluster associated to him/her. Subsequently, the skill-matching system will provide the same jobs ( i.e A,B,C from *Figure 7*) that were associated to previous users or migrants of the same cluster.

Consequently, the system is more robust and complete by taking in consideration the similar users that are in the platform and who have similar competences.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

## 3.4 Skill-Matching Implementation Design

As mentioned in the previous sections, the module of skill matching depends on the information from the mathematical embeddings of T3.1. Skill matching module will rely on the outputs of such a module, therefore it will be directly connected, and also the recommendation engine (T3.3) can use information based on inferences from this module.

For this purpose, the initial approach involves the following technologies:

- **Flask:** which is a WSGI web application framework based on python. This component will provide interoperability of the Skills matching module with the rest of ReBUILD tools.
- **Neo4j: i**s a graph database management system with native graph storage and processing.
- **Python:** will be employed as the programming language for the algorithms application and connectivity between Flask and Neo4J. Related libraries such as PyNeo[6].

Furthermore, these components will be contained in a docker virtualized environment to facilitate the Continuous Development of ReBUILD platform.

In this module, a dockerfile will be created for the neo4j and environment creation with a volume for the internal graph database. Additionally, this docker will expose some ports for interaction. In concrete, the port 5000 for the API connection, and ports 7474 and 7678 for the Neo4j web browser. Additionally, the scalability is reached by the separation of the volume, permitting fast response and adaptability to a large number of users; modularity of these components is guaranteed due to the fact that these components will be integrated into a centralized platform so the images will be uploaded as Open in Docker Hub repository, as well as documentation for allowing skills matching research.

Finally, *Figure 8* represents the general architecture of this component in terms of software design using the aforementioned technology.

[6] PyNeo library available at https://py2neo.org/v4/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

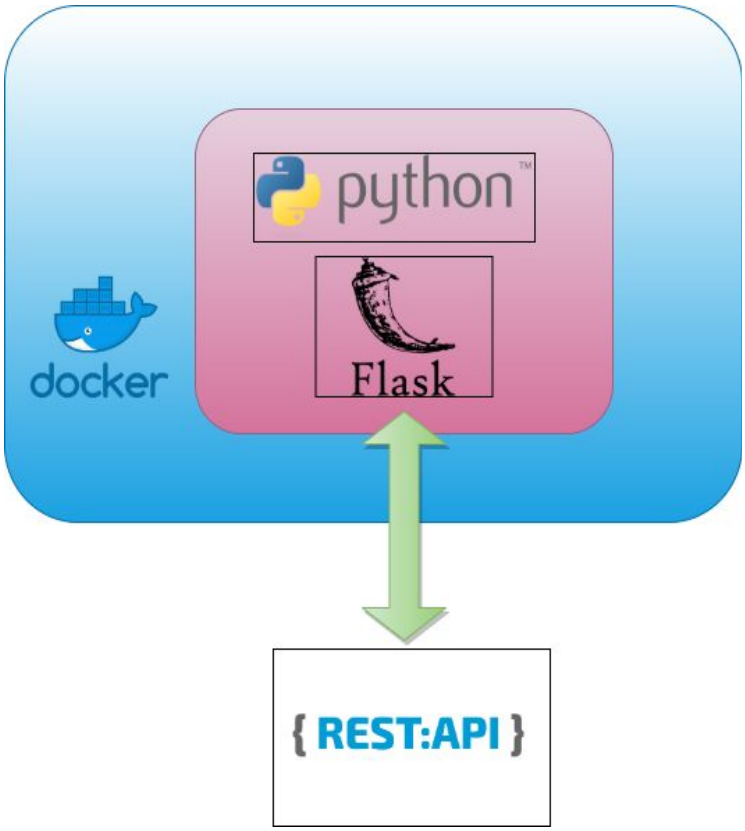


*Figure 8: Overall architecture of the Skill-matching in terms of software design and implementation .*

# 4 CONCLUSION

This deliverable is a prototype regarding the skill-matching component of the REBUILD project. The goal of this component is to provide useful information for migrants across different topics such as job seekings where the most adequate jobs for a given candidate are provided according to their competences. Thus, the system attempts to associate latent information, i.e skills in the job seeking scenario, together with a target objective i.e jobs in the job seeking scenario.

Moreover, the skill-matching system can be generalized and extended to other use cases as long as a clear definition of both the latent and the target information are provided and available in the data model.

Along this document, the main techniques employed within the skill-matching framework were described including description logic, ontology-based, machine learning and hybrid approaches. More specifically, the hybrid solution was declared as the most robust and adequate for the purposes of the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822215

project, considering that the system will receive as input a set of user embeddings from Task 3.1 which represent the user profiles in a lower space.

Furthermore, the deliverable describes different ontologies that were analysed for researching purposes including the so-called ESCO dataset and a customized ontology for the REBUILD project focussed on job seeking recommendations. A graph framework named NEO4J was employed to run all the experiments in order to verify its utility within this task. In particular, it will be used to gather the different elements or entities that are needed on each service to build a graph where all the elements are related according to some relationships defined in the REBUILD data model. Finally, on top of this graph structure, several algorithms and techniques can be calculated with the aim of assessing a higher precision in the skill-matching task.

It is important to remark that some of the definitions of the entities or relationships of the presented ontology may change with the integration of new services and functionalities as well as due to the restrictions that the GDPR law may have. However, this document will be updated in a further version, where the final version of the skill-matching prototype will be detailed.

Finally, this document presents an initial design of the architecture for the skill-matching component using virtual containers such as Docker as well as a Python web-based framework named as Flask. Hence, the system will have its own RestFul-API in order to facilitate the communication with other components.

The upcoming work will address several open issues: (1) The integration in the ReBUILD platform according to the integration plan WP5. The second branch (2) is mainly focused on the data collection, analysis and optimisation of the existing models. Finally, the third aspect is related to the use of other techniques in the SoA to be compared with the results presented here.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

# 5 REFERENCES

Petrican, T., Stan, C., Antal, M., Salomie, I., Cioara, T., & Anghel, I. (2017, September). Ontology-based skill matching algorithms. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 205-211). IEEE.

Calì, A., Calvanese, D., Colucci, S., Di Noia, T., & Donini, F. M. (2004, September). A logic-based approach for matching user profiles. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 187-195). Springer, Berlin, Heidelberg.

Bizer, C., Heese, R., Mochol, M., Oldakowski, R., Tolksdorf, R., & Eckstein, R. (2005). The impact of semantic web technologies on job recruitment processes. In *Wirtschaftsinformatik 2005* (pp. 1367-1381). Physica, Heidelberg.

Hohaghegh, S., & Razzazi, M. R. (2004, June). An ontology driven matchmaking process. In *Proceedings World Automation Congress, 2004.* (Vol. 16, pp. 248-253). IEEE.

Hassan, F. M., Ghani, I., Faheem, M., & Hajji, A. A. (2012). Ontology matching approaches for erecruitment. *International Journal of Computer Applications*, *51*(2).

Botov, D., Klenin, J., Melnikov, A., Dmitrin, Y., Nikolaev, I., & Vinel, M. (2019, June). Mining Labor Market Requirements Using Distributional Semantic Models and Deep Learning. In *International Conference on Business Information Systems* (pp. 177-190). Springer, Cham.

Fazel-Zarandi, M., & Fox, M. S. (2009). Semantic matchmaking for job recruitment: an ontology-based hybrid approach. In *Proceedings of the 8th International Semantic Web Conference* (Vol. 525).

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2004). Ontology matching: A machine learning approach. In *Handbook on ontologies* (pp. 385-403). Springer, Berlin, Heidelberg.

Sayfullina, L., Malmi, E., & Kannala, J. (2018, July). Learning representations for soft skill matching. In *International conference on analysis of images, social networks and texts* (pp. 141-152). Springer, Cham.

Malinowski, J., Keim, T., Wendt, O., & Weitzel, T. (2006, January). Matching people and jobs: A bilateral recommendation approach. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)* (Vol. 6, pp. 137c-137c). IEEE.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215

REBUILD

ICT-enabled integration facilitator and life rebuilding guidance

# Deliverable: D3.2 Methodology for skills and needs matching

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 822215.