DISCOVERING JOB KNOWLEDGE FROM WEB DATA

Intellectual Output 2

Online Data Sources for the European Labor Market

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Disclaimer: The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the European Union

Erasmus+
Enriching lives, opening minds.

DISCOVERING JOB KNOWLEDGE FROM WEB DATA

IO2 Report Title: Online Data Sources for the European Labor Market

Project Title: Discovering Job Knowledge through Web Analytics towards facilitated mobility of European Professionals and Refugees Career Integration

Project Coordinator

Institution: GOTTFRIED WILHELM LEIBNIZ UNIVERSITAET HANNOVER

Address: Welfengarten 1, 30167 Hannover, Niedersachsen, De

Website: www.uni-hannover.de

Report Prepared By: PETANUX GMBH, Bornheimer Straße 127, 53119 Bonn

Project Information

Identifier : 2018-1-DE02-KA202-005215

Start End : 01-SEP-2018

End Date : 31-AUG-2021

Partners : ENGINEERING - INGEGNERIA INFORMATICA SPA (IT)
EKONOMICKY USTAV SLOVENSKEJ AKADEMIE VIED (SK)
PETANUX GMBH (DE)

Website : https://www.diskow.eu/

Report Version Details

Revision # 01 : Initial Internal Release on Feb 14, 2020

Revision # 02 : Updated Internal Release on Feb 21, 2020

Revision # 03 : Updated Internal Release on Feb 27, 2020

Revision # 04 : Final Release on Mar 19, 2020

Table of Contents

Executive Summary 3
List of Abbreviations and Acronyms 4
1. Introduction 5
2. Tasks Constituting IO2 6
2.1 O2/A1 - Identification and classification of JK Data Sources from the Web 6
2.2 O2/A2 - Evaluation and Verification of open standards for JKB implementation 6
3. Job Portals: Global, European & Local 7
4. Country-wise Job Portals and Job Market Scenario 8
4.1 Slovakia 8
4.2 Germany 11
4.3 Italy 13
5. Shortcomings of Existing Job Portals 15
6. Potential Data Sources for DISKOW Project 16
6.1 EURAXESS 16
6.2 EURES 16
6.3 DBWorld 16
6.4 CareerJET 17
6.5 Profesia 17
6.6 Indeed 17
7. Semi-structured Data vis-à-vis Unstructured Social Media Data 18
8. Comparison of Proposed Schema of IO1 with ISCO classification 21
9. Comparison of Proposed Schema with Existing Job Portals 24
10. Conclusion 26

Executive Summary

IO2 report is broadly addressing two main tasks of the DISKOW project: identification of potential data sources and evaluating them with the proposed schema of IO1. Hence, the first part of the report mostly focused on the identification and classification of different online data sources. This report has classified the job portals into three categories: global job portals, European job portals, and local country-specific job portals. This report has explored multiple existing job portals. Next, this report has tried to identify the most relevant and accessible data sources for developing JKB. Subsequently, DISKOW project partners are contacting these job portals for accessing the data. In addition, this report has also explored open data sources such as social media. This report has explored Twitter data as one of the potential sources for social media data and discussed the merits and demerits of this open data source. This report has also developed a framework to create a job knowledge base from social media data. The next portion of the report extends the findings of IO1. This report has tried to evaluate and verified the schema prepared by IO1 in the context of various potential data sources, and finally, this report has suggested how the data sources, identified by IO2, and the schema prepared by IO1, can be integrated into the subsequent intellectual outputs.

List of Abbreviations and Acronyms

Sl. No.	Abbreviation	Description
1	API	Application Programming Interface
2	EQF	European Qualifications Framework
3	ESCO	European Skills, Competences, Qualifications, and Occupations
4	EURES	European Job Mobility Portal
5	ILO	International Labour Organisation
6	ISCED	International Standard Classification of Education
7	ISCO	International Standard Classification of Occupations
8	ISCO-08	International Standard Classification of Occupations – version 2008
9	ISCO-88	International Standard Classification of Occupations – version 1988
10	JK	Job knowledge
11	JKB	Job knowledge base
12	JSON	JavaScript Object Notation
13	LOD	Linked Open Data
14	OWL	Ontology Web Language
15	XML	Extensible Markup Language
16	XSD	XML Schema Definition

1. Introduction

Unemployment is a serious global problem in recent times. Effective dissemination of job-related information to job seekers could partially address this unemployment issue. Job-related information can be of two types: what all jobs are available? And, what job providers are mentioning for a specific type of job? Hence, this DISKOW project aims to develop a Job Knowledge Base (JKB) by collecting and analyzing job-specific data from various online sources. This JKB will help the job seekers to understand the required competencies and skills for a specific job. Leibniz Universität Hannover (LUH), Germany, is coordinating this DISKOW project.

This second section of the DISKOW report, i.e., Intellectual Output 2 (IO2), is exploring publicly available web-based information to understand the dynamics of the European Job Market. This report has studied various types of job portals, especially in the European context, according to the scope of this Erasmus+ DISKOW project. Petanux GmbH, Bonn, Germany, has prepared this Intellectual Output 2 (IO2). To prepare this report, Petanux GmbH has consulted project partners from other countries for their country-specific inputs like challenges in the local labor market or popular portals in these respective countries. Two key partners of the DISKOW project are Ekonomicky Ustav Slovenskej Akademie Vied from Slovakia and Engineering - Ingegneria Informatica Spa from Italy. So, two target countries, in addition to Germany, which this report has studied, are Slovakia and Italy.

This IO2 report explores multiple commercial job portals and the nature of their services to develop the Job Knowledge Base (JKB). It was observed that most of the job portals are offering free information up to a certain limit to both individual job seekers as well as corporate job providers. However, to get customized service, corporate users, as well as individual users, need to pay a service charge. Mostly, this service charge depends on the nature of services and the duration of the service. For instance, how many months or weeks a corporate job provider wants to use the dedicated service form a job portal.

As it was mentioned earlier, this report has considered three countries, where DISKOW partners have first-hand experience and local knowledge, namely Germany, Italy, and Slovakia. This report has identified leading job portals in these countries and explored the nature of job-related information provided by these portals. Next, this report has carefully analyzed them and made some comparative analyses. This report has also considered a few global job portals like LinkedIn, Glassdoor, in addition to local job portals, which are offering customized services in these countries. According to the DISKOW Grant proposal, this report has confined itself in the domain of Computer Science, Information Technology, Web Development, and Data Science related jobs. However, this study could be extended to other contexts.

It is important to note that there are certain limitations of existing job portals, which is also stated in the legal notice of these portals. For instance, according to the terms and conditions of one job portal - “You can use the material available on this site for any non-commercial purpose. Usage of material for a commercial purpose is not allowed.” This type of declaration is common nowadays to ensure data privacy and protect their proprietary data. Most job portals

charge a fee, according to their service policy, to provide the customized service. Customized services can be of two types: data of job seekers to job providers and job provider postings to job seekers. Hence, if this project plans to use the data from commercial job portals, then DISKOW partners might have to make an official (or legal) agreements with these job portals for using their data to develop the JKB.

To address the promised tasks of IO2, this report has extracted data in the following manner. Initially, by browsing the websites of existing job portals, this report has extracted the freely available data from these paid websites. As it was mentioned earlier, most of these job portals are paid portals for customized service. Hence, this report has also identified data sources where job portals allow to legally access their data. In addition, this report also offers some insights regarding alternative free data sources such as Twitter. Next, this report has evaluated these job portals on the basis of the proposed schema of IO1. In brief, this report explores the European job market scenario, specifically in the domain of computer science and information technology-related jobs, and evaluates various data sources, i.e., exiting job portals, using the schema proposed by IO1.

2. Tasks Constituting IO2

This report addresses the two main tasks that constitute the IO2, following the Description of Work (DoW) in the DISKOW Grant Agreement. These tasks are reproduced below from the DISKOW proposal, and subsequent portions of the present IO2 report has addressed them rigorously.

2.1 O2/A1 - Identification and classification of JK Data Sources from the Web

Different data sources will be used to identify the JK from the web, and proper legal agreements to access available data will be established with data owners. In this regard, more emphasis will be given to the sources which are potentially more useful in this regard. Open data sources such as social media, job advertisement and position announcements, as well as other data sets such as EC provided data sets will be considered and the exploitation strategies are devised. This task will also consider the classification of the available data sources based on the ontology done in IO1.

2.2 O2/A2 - Evaluation and Verification of open standards for JKB implementation

Open standards such as HR-XML and Data Model for Reusable Competency Definitions (IEEE 1484.20.1) will be covered in this task and a new standard that fulfills the advantages of those standards and targets innovations in the DISKOW project will be developed. As a result, by using proposed standards in this task, the JKB will be able to be integrated into different HR software systems.

3. Job Portals: Global, European & Local

The following section probes various existing job portals and services they provide. We have classified the existing job portals into three categories (refer to Figure 1). First, global job portals which have presence across the world. Second, some of the major European job portals which have its presence across European countries.

Third, we have also tried to identify some local country-specific job portals which are popular within the boundary of a country. Hence, we have identified the major job portals in Germany, Slovakia, and Italy. Also, this section provides a brief description of these job portals. It is important to note that some of the job portals have a global presence, but they offer country-specific services in these countries. For example, Monster is a global job portal but job seekers in Germany mostly use their Monster Germany services. Similarly, LinkedIn has global presence but in Italy job seekers and job providers mostly explore 'LinkedIn Italy'. Hence, in this report, we have tried to explore their country-specific services – if it is available. Otherwise, we have considered their global job portals and restricted our search by mentioning the geographic region of job search. In short, in this section, we have probed each of these portals and extracted data for available jobs in each of these countries.

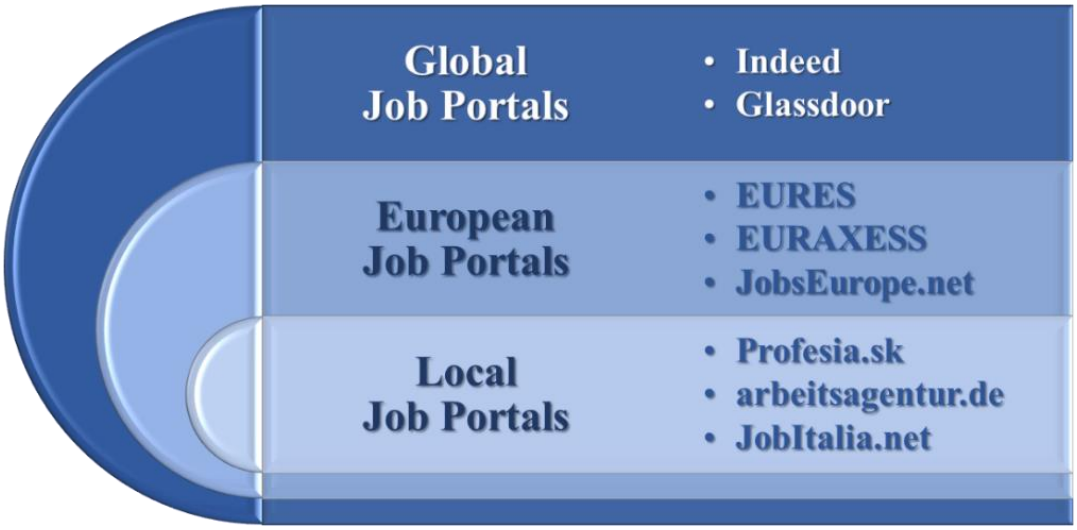

Figure 1:Three Levels of Job Portals

As we have mentioned earlier, we have extracted job-related data only in the domain of 'computer science' and 'information technology'. However, these are generic phrases and search results might not be very conclusive. Hence, we have also considered two specific sub-categories namely, data science and web development. To maintain parity with IO1, we have focused on Data Science related jobs in our all three target countries.

4. Country-wise Job Portals and Job Market Scenario

4.1 Slovakia

We have identified various existing job portals in the context of Slovakia, as listed in Table 1, and we have also explored the job postings in these portals – as reported in Table 2. The case of Slovak labour market is an example of dominance of one online job portal and missing reliable ground truth statistics for vacancies. Since 2006 private online job portal "Profesia.sk" holds from 6 to 8 times more job advertisements than is in the official evidence of Slovak authorities[1] [2]. Kureková et al. (2015)[2] claims, that if in countries where a dominant portal exists, collected vacancies might be the best available source. Profesia is a technology company based in Slovakia. They mostly focus on the development of software products and services for the human resource management, labor market, and recruitment system like job portal. Table 2 clearly indicates that the number of jobs posted in Profesia is much higher than all other job portals in the context of Slovakia. We have contacted our Slovakian partners in this DISKOW project for their feedback, and they have also confirmed that Profesia is the leading job portal in Slovakia.

Our Slovakian partner has already contacted Profesia for accessing their database to develop the JKB and made an agreement with profesia.sk for the DISKOW project. This dataset will help us to develop a better JKB. Our subsequent intellectual outputs, i.e., IO3, will report the details of the Profesia dataset and findings from the same in an elaborate way. In addition to Profesia database, this IO2 report has also probed other dominant job portals in Slovakia and explored the job postings in these portals to get a holistic view of Slovakian job market. As we mentioned earlier, these job portals are not so dominant because the number of job postings are significantly lower with respect to Profesia. So, our subsequent intellectual outputs will not consider these portals.

Sl. No.	Job Portals	Description
1	www.profesia.sk	A leading job portal in Slovakia.
2	www.jobs.cz	It is worth noting that the domain "cz" is for Czechia. However, this Czech portal covers common Czech and Slovak labour market including also job advertisement from Austria and Germany. This portal offers jobs to Engineers, Academicians, Finance, and Marketing professionals not only for freshers but also for experienced job seekers.

[1] Miroslav Beblav`y, Lucia M`ytna Kureková, and Corina Haita. 2016. The surprisingly exclusive nature of medium-and low-skilled jobs: Evidence from a Slovak job portal. Personnel Review 45, 2 (2016), 255–273.

[2] Lucia M`ytna Kureková, Miroslav Beblav`y, and Anna Thum-Thysen. 2015. Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. IZA Journal of Labor Economics 4, 1 (2015), 18.

3	www.job.sk	*Job.sk* is a portal that offers jobs to employee seekers in different regions in Slovakia and Czechia. However, this portal mostly provides jobs in Slovakia. For instance, they have around 350 new jobs in Slovakia and only 15 jobs in Czechia.
5	www.jobagent.sk	*Jobagent.sk* is a small job portal that has listed around 250 to 300 jobs in the Slovak region. They have started operating in the year 2005. However, the growth of this portal is not very impressive. They have categorized their job according to the position, industry, and location.

Table 1: Job Portals from Slovakia

Source: Data/information about these portals were extracted from their respective websites.

Next, figures 2 and 3 provide screenshots of two job portals to highlight the nature of the user interface. Most portals allows searching jobs based on location, designation and industry. For instance, www.Jobs.cz also allows for adding multiple filters. Job portals, such as www.profesia.sk, report location, salary and detailed job description in their site. Most of these portals are user-friendly for job seekers.

Sl. No.	Domain	profesia.sk	jobs.cz	www.job.sk	jobagent.sk	European Job Mobility Portal
1	Computer Science	164	724		40	38
2	Information Technology	2449		57	13	111
3	Data Science	112	-	-	-	514
4	Web Development	229	NA	-	-	365

Table 2: Job Market Data in the context of Slovakia

Source: Above data-points were extracted during Feb 2020 from their respective websites

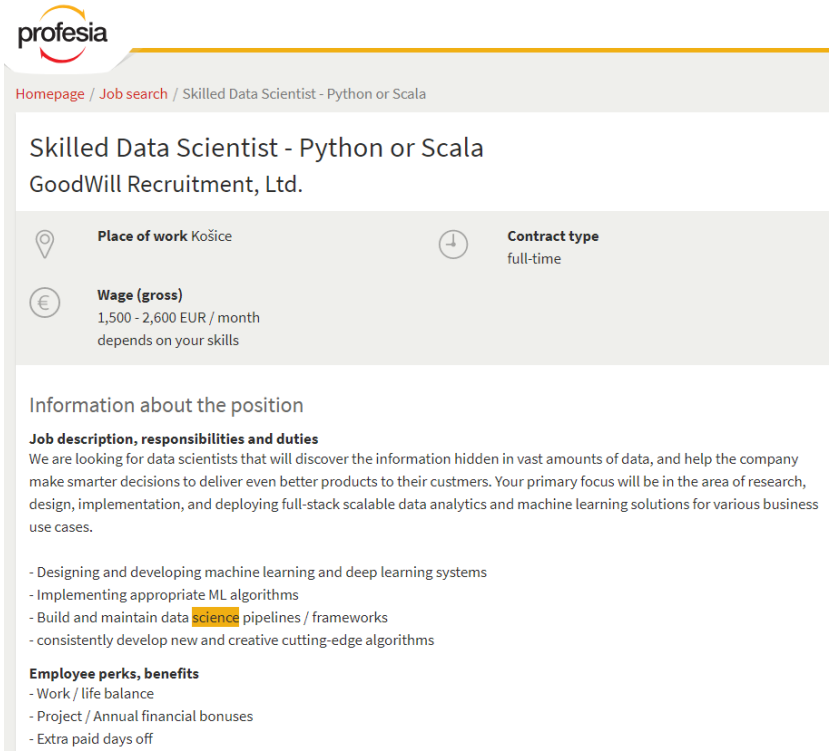

Figure 2:Sample Data Scientist Job description @ www.profesia.sk (Source: Portal website)

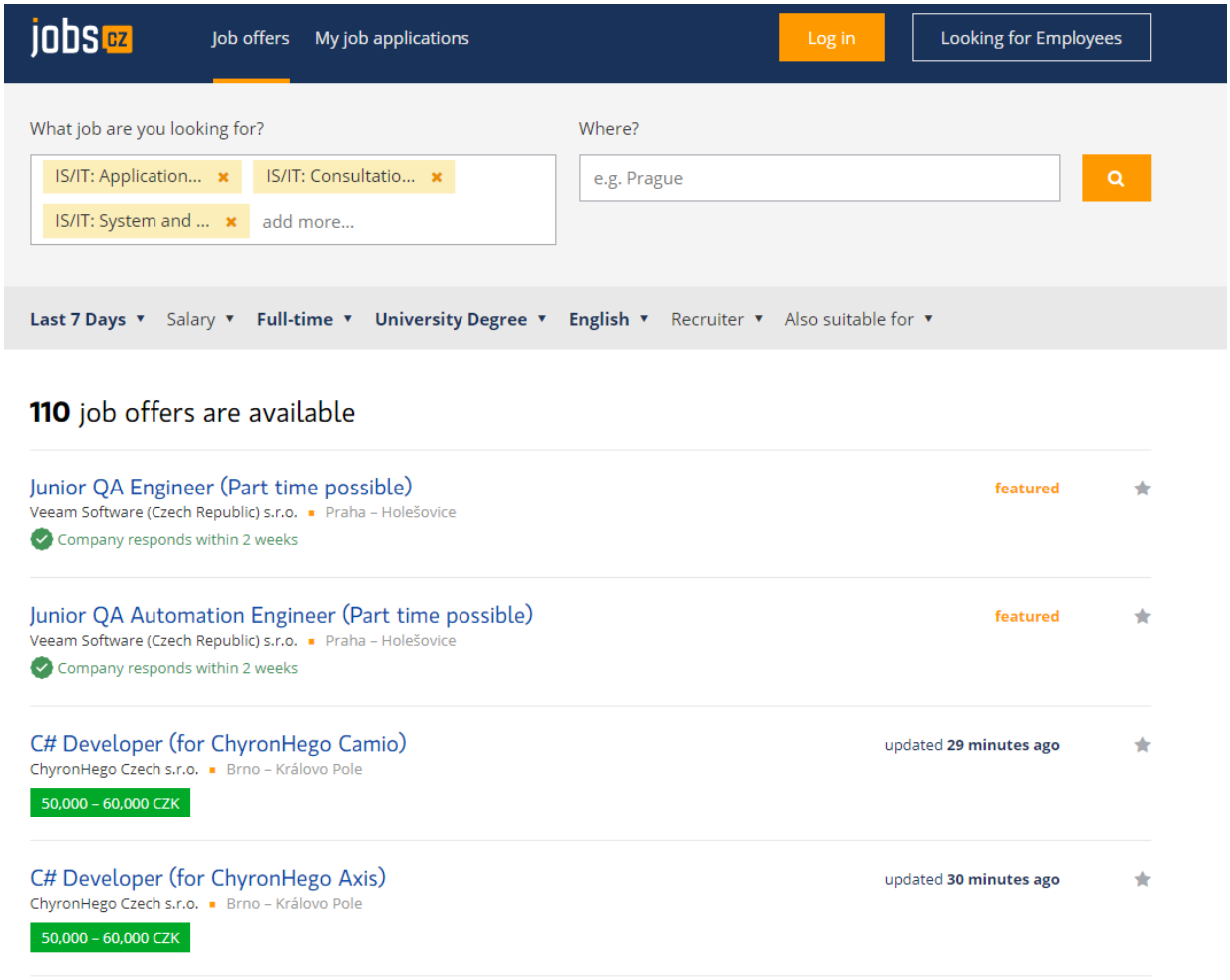

Figure 3: Search results with different filters like Degree, Working language, Job posting time and Job type @ *www.Jobs.cz (Source: Portal Website)*

4.2 Germany

Table 3 is reporting the major job portals used by job seekers in the German Context. Similar to Slovakia, Table 4 reports the computer science and information technology job-related data in the context of Germany. Similar to the previous country, the number of listed jobs for generic phrases, such as Computer Science or Information Technology, is much higher than specific 'Data Science' jobs. However, we are still in the process of identifying the most suitable job portals in the German context, and after identifying the same, we have to make an agreement with them for accessing the data for developing the JKB. So, this IO2 report provides a holistic view of all major job portals in the German context.

Sl. No.	Job Site Name	Description
1	Monster Germany	This portal is mostly used by the German job seekers to search German jobs. The Monster started in 1994, is one of the oldest web services. They offer job recommendation service. It is important to note that Monster offer this service all over the world.
2	StepStone Germany	Stepstone provides job-related information across various categories/domains for young professionals as well as experienced job seekers. They also provide paid services to corporate houses as well as to individual job seekers.
3	Jobtopus	Jobtopus is a German job portal that offers services mostly in the regions of Frankfurt, Cologne, Stuttgart, Berlin, Dusseldorf, Hamburg, Hanover and Munich.
5	Indeed Germany	Indeed.de is one of the largest job portals. For instance, it provides thousands of jobs in each of the sub-domains of computer science like web development, data science. They have structurally categorized the data for various types of jobs.

Table 3: Job Portals from Germany

Source: Data/information about these portals were extracted from their respective websites.

Sl. No.	Domain	Monster Germany	StepStone Germany	Jobtopus	Indeed.de	European Job Mobility Portal
1	Computer Science	14844	12993	13037	5970	19,027
2	Information Technology	17615	15373	28391	6737	432178
3	Data Science	347	5899	5932	6164	8420
4	Web Development	2803	8360	8503	7155	68785

Table 4:Job Market Data in the context of Germany

Source: Above data-points were extracted during Feb 2020 from their respective websites

4.3 Italy

Table 5 has listed the major job portals and Table 6 is reporting computer science and information technology job-related data in the context of Italy. Our Italian partner has identified some of leading job portals of their country and they are in the process of making an agreement with them. So, in this report, we are providing a holistic view of the major Italian job portals and the volume of job postings in these portals. Interestingly, the number of jobs in the Data Science category and the number of jobs in the domain of Computer Science and Information Technology are not drastically different. We have looked into it carefully and it seems that job postings are mostly using specific job requirements and not using generic phrases. We need to probe this further. Figure 5 provides the screenshot of computer science-related job postings in the EURES website to highlight its user-friendly interface.

Sl. No.	Job Site Name	Description
1	Indeed Italy	Indeed Italy offers a range of jobs across the different sectors. According to their website, the total number of jobs in Italy in February 2020 is 103,364. Italian Job Seekers and Employers frequently use Indeed Italy.
2	Jobs in Milan	It is a Milan-based company for the Italian labor market. In addition to the Italian labor market, it also lists a few jobs offers beyond Italy – such as Germany and France. Their portal also allows users searching using multiple languages such as English, German, French in addition to the Italian language.
3	CareerJet Italy	Career Jet Italy is another Italian job portal. Here, job seekers need to properly fill the search query in advanced mode for extracting specific job details.
4	Linkedin Italy	LinkedIn is another global joint in this domain. LinkedIn Italy offers customized services for the Italian market.
5	jobitalia.net	Job Italia is a fully Italian joint-stock company. Their consultancy service aimed at identifying suitable candidates to fill one or more job positions.

Table 5: Job Portals from Italy

Source: Data/information about these portals were extracted from their respective websites.

Sl. No.	Domain	Indeed Italy	Glassdoor Italy	JobsinMilan	CareerJet Italy	Linkedin Italy	European Job Mobility Portal
1	Computer Science	1,214	1225	5812	658	1000+	8
2	Information Technology	2.839	2069	7,849	5940	2000+	10
3	Data Science	1.179	304	1,290	739	976	16
4	Web Development	965	245	1,708	823	770	11

Table 6:Job Market Data in the context of Italy

Source: Above data-points were extracted during Feb 2020 from their respective websites

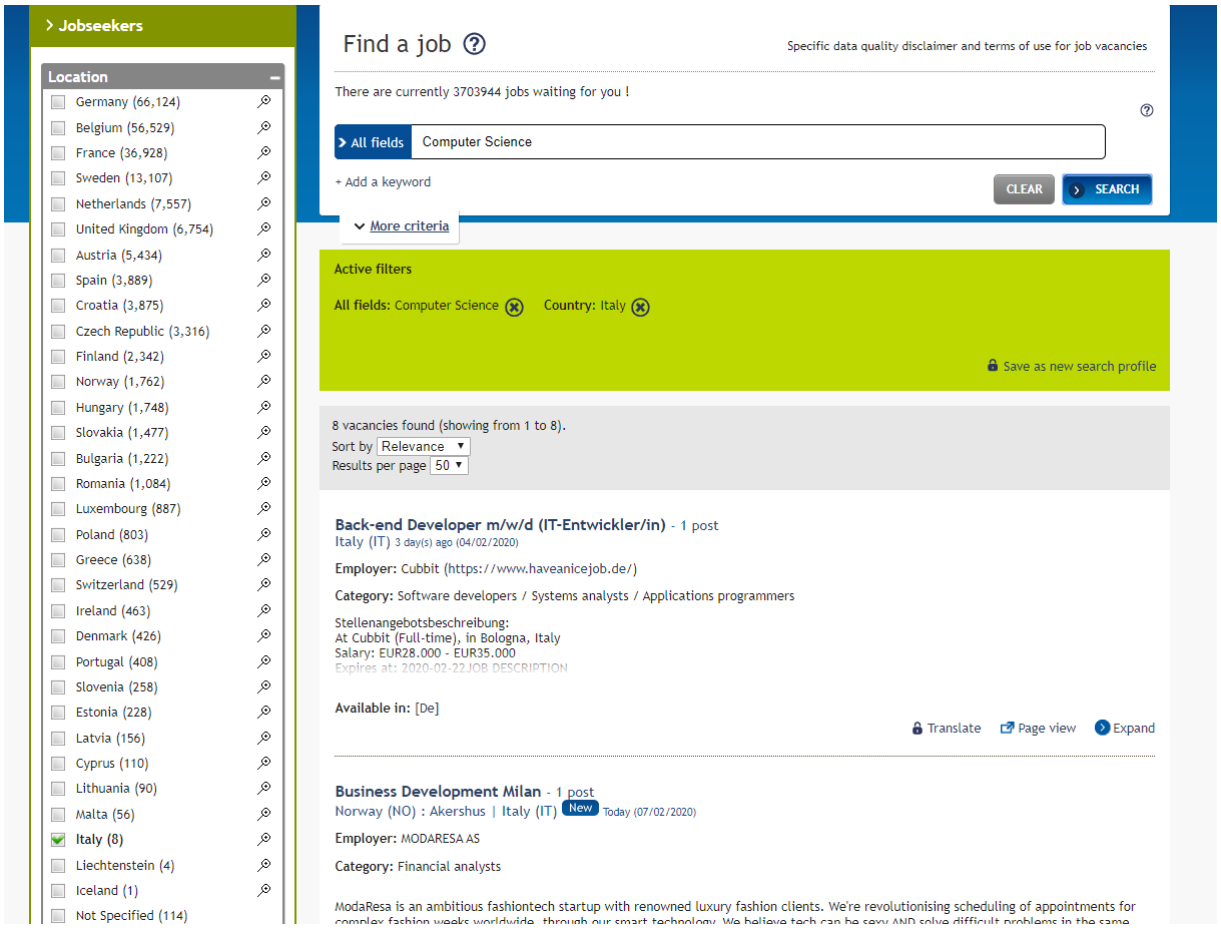

Figure 4:Search result for Computer Science in EURES website *(Source: Portal Website)*

5. Shortcomings of Existing Job Portals

In the previous section, we have extracted job-related data from various existing job portals. We have extracted data for four types of jobs. As we mentioned earlier, that some job postings are using generic phrases whereas other job postings are using specific terms. As a result, our search outputs might overestimate (or underestimate) the available jobs. This is a potential problem of the existing job portals. To highlight these problems, Table 7 reports our searching methodology in tabular format. Here, it is important to note that search 'key' words, such as 'Computer Science' or 'Information Technology,' are generic phrases and can be used across different types of computer science-related jobs. On the contrary, 'Data Science' is for a specific type of Computer Science job. For example, the first two job postings (Sl. No. 1 & Sl. No. 2) are exactly same but according to our search it is coming under both 'Computer Science' category as well as 'Data Science' category. However, this is not the case for last two job postings (Sl. No. 6 & Sl. No. 7). So, these portals might give accurate results for a specific category but interpreting the outputs for generic phrases will be problematic. Using multiple phrases can also lead to double counting. Hence, a structured approach, such as ISCO-08 coding or HR-XML or the proposed schema of IO1 can addressed this problem.

Sl. No.	Posting	Search 'Key' Word	Search Results
1	Hiring a Computer Science Engineer with prior experience in Data Science	Computer Science	Yes
2	Hiring a Computer Science Engineer with prior experience in Data Science	Data Science	Yes
3	Hiring a Computer Science Engineer with prior experience in Web Development	Computer Science	Yes
4	Hiring a Computer Science Engineer with prior experience in Web Development	Data Science	No
5	Hiring a Computer Science Engineer with prior experience in Web Development	Web Development	Yes
6	Hiring a senior-level engineer with prior experience in Data Science	Computer Science	No
7	Hiring a senior-level engineer with prior experience in Data Science	Data Science	Yes

Table 7: Our Searching Methodology and Outcome

In addition, some of these postings mention the required skills sets whereas some of them don't mention. For example, interpretation of 'prior experience' might vary from one job provider to another job provider. Also, many of these job postings don't mention 'job location' or 'salary'. Table 10, in the later portion of this DISKOW report, provides a detailed analysis. In short, the nature of job postings (and missing information) justifies the need to develop a proper data model.

6. Potential Data Sources for DISKOW Project

We have identified multiple sources for legally accessing the data for European Job Market. The following section reports the legal disclaimer/notification provided on their website. These disclaimers clearly mention that this data can be freely accessed. So, DISKOW project partners can officially contact them to pursue these sources further and explore the possibility of using these sources for developing our proposed JKB.

6.1 EURAXESS

This is a potential data source for European job market because their website allows to download all jobs in XML format. The website also gives the usual disclaimer that "The responsibility for the jobs published on this website, including the job description, lies entirely with the publishing institutions. The application is handled uniquely by the employer, who is also fully responsible for the recruitment and selection processes." However, this is true for any job portal. For example, neither Indeed nor Glassdoor will take the responsibility of the job postings.
Website: https://euraxess.ec.europa.eu/jobs/search
Contact Details: support@euraxess.org

6.2 EURES

This is another potential data source for the European job market. EURES is a European Commission job mobility portal, and more importantly this "can be accessed free of charge by both jobseekers and employers. Jobs advertised on the EURES portal come from EURES members and partners, in particular the European Public Employment Services." In the context of DISKOW project, we found that "only EURES member and partner organizations, recognized by a EURES National Coordination Office, are allowed to extract data using our API or similar technologies. In this case the European Commission cannot give access to data, owned by a third party such as employers for another use than the one in the context of the EURES portal."
Website: https://ec.europa.eu/eures/public/en/help-and-support
Contact Details for EURES partnership/membership:
https://ec.europa.eu/eures/public/en/how-to-become-a-eures-partner/member

6.3 DBWorld

This site allows to read and post messages using their web interface. The DBWorld website says that "Anyone can browse messages, but you must register to obtain a DBWorld posting password before you can post messages through the Web. Posted messages are sent immediately to the dbworld list without being moderated by the list owner." So, in the initial stages they don't monitor the posts by users, but they also mentioned that "Posting messages that are not appropriate to this list can lead to loss of posting privileges, so please avoid posting irrelevant messages."
It is important to note that this database is a very domain-specific portal mostly for computer science professionals. Job postings offer positions such as research assistant,

research associate, PhD position, Post-doctoral positions, faculty positions and other academic/research related jobs. In addition, this website also offers information about upcoming conferences and journal call for papers.
Website: https://research.cs.wisc.edu/dbworld/
Contact Details: ballard@cs.wisc.edu and anhai@cs.wisc.edu
We are thinking of two options for accessing this database. Either we will coordinate with the contact persons mentioned in their website or we will develop a customized web scrapper for scrapping the specific job announcement. As we have mentioned earlier, this website not only mentions job postings but also other announcements, such as conference related information. Hence, our web scrapper should need to identify and scrap job postings data for developing the JKB for the DISKOW project. In the next intellectual output (IO3), we will address this.

In addition to the above data sources, there are few job portals which allows API support to extract data. In the later stages of DISKOW project we might explore them. The following portion provides details of two such job portals which can be explored in IO3.

6.4 CareerJET

Careerjet is a job search engine that fetches job postings from different portals/websites like academic institutions, private companies as well as government organizations. Most importantly, they also provide API support for third-party users. This API is compatible with PHP versions 4 and 5 according to the documentation/information shared in the CareerJet website. The CareerJet API allows developers to embed CareerJet search results into their websites. Search can be performed based on keywords, locations, companies as well as industries. So, this is another potential data source which can be explored later.
Website: https://www.careerjet.com/
Contact Details: https://www.careerjet.com/about/contact.html

6.5 Profesia

As we have mentioned earlier, one of our project partners is coordinating with them for accessing their database.

6.6 Indeed

Similar to CareerJET, Indeed job portal also provides different types of API for third-party services like Job Search API, Indeed Apply, XML Job Feed etc. Using the Job Search API of Indeed, anyone can submit new job searches and export the search results on their websites. However, they have not provided detailed information about accessing their database by using API, but this can be explored in future.
Website: https://www.indeed.jobs/
Contact/registration Details: https://www.indeed.jobs/career/Login

7. Semi-structured Data vis-à-vis Unstructured Social Media Data

As we mentioned earlier, conventional commercial job portals, such as LinkedIn or CareerBuilder, are providing semi-structured data of job providers and job seekers. Exiting research has used this semi-structured data for job recommendation tasks, but most of these studies are the outcome of their in-house research. So, the problems of these portals are two folds. First, these semi-structured data in these job portals are proprietary; thus, they are not sharing their data publicly. Secondly, mostly these portals charge their users for customized service. So, big corporations or rich jobseekers can afford their customized services, but small organizations and not-so-rich job seekers might not be able to afford it.

Hence, we need to probe alternate sources of data. Twitter is one such alternative sources of data. Firstly, it is a freely available data source, and prior research used these data for addressing various socio-economic issues like social movements, disaster management, stock market behaviors etc. However, Twitter data might not be representative of the society and the data is unstructured. So, the question is: how jobseekers will get the relevant information from the Twitter platform. We find that most organizations, even if they are small, are using Twitter platforms for promoting their brand and reaching customers. Nowadays most recruiters are present on social media platforms. More importantly, they also use these social media platforms for sharing vacancies, new jobs, and recruitment plans on social media platform. Twitter is a popular communication channel not only for big organizations but also for small organizations, who cannot afford traditional job portals, to reach the labor market. This trend will be more popular in the coming days. Table 8 reports some sample job-related tweets and attributes through color-coding.

We suggest crawling this publicly available social media data and to use various NLP techniques to extract relevant information from this unstructured data. Using Twitter Search API, we can extract job postings from Twitter platform without any legal issues which is a major limitation for most commercial job portals. We can extract job-related tweets from Twitter API[3] using different job-related keywords like: ‘job’, ‘vacancy’ etc.

As a pilot study, we have collected some job-related tweets and analyzed their content. We have tried to explore whether Twitter job postings are satisfying the proposed schema of IO1. In Table 8, we analyzed the content and tried to identify various attributes of job-related tweets. We found that many job-related tweets clearly describe the expectations from job seekers. In addition, these tweets also mention - What is the overall scope of that particular job? Which location? What skill sets are required? However, it is important to note that all job-related tweets are not so informative. Broadly, we conclude that job postings on social media platform can be used as a potential data source in the coming days, but to efficiently use it – we need to have a structured approach.

In previous section, we have mentioned various shortcomings of commercial job portals. In addition to those shortcomings, another challenge of existing commercial job portal is lack of

[3] https://developer.twitter.com/en/docs/tweets/search/api-reference

variations. In other words, these portals lack heterogeneity in terms of variety of job postings. For example, one job portal might be known for only finance-related jobs in one particular location/country. As a result, job providers in IT domain might not opt for that particular portal. However, this is not the problem for social media data. In other words, by using social media data, which comes from all over the world, we can overcome this lack of heterogeneity in terms of domain or location.

Sl. No.	Sample Job Posting-related Tweets	Location	Job Description	Skills	Salary
1	With SAS analysis experience, an opportunity to join a lovely agency in SW London in this 9 – 12-month Mat Cover Data Analyst role. #marketingjobs #newjob #dataanalyst #analytics #marketinganalyst #SAS #SASprogramme	Y	Y	Y	
2	JOB; Dallas TX USA - Inside Sales Representative - The B2B Inside Salesperson is responsible for making outbound calls to management staff at prospect companies. Your focu JOBS WORK #DALLAS	Y	Y	Y	
3	Jobs Application Support Analyst, Chippenham Application Support Analyst What you need to have At least 2 year's IT support experience in a Application Support Analyst	Y	Y	Y	
4	Are you interested in designing #fullstack #code for web applications in the financial industry? Do you have experience using #Angular, #typescript, #SQS, #Nodejs or #Oauth? Click here #careers #WisconsinJobs #JavaDevelopment #Engineer #JavaScript #HTML5	Y	Y	Y	
5	iOS Developer Johannesburg, Gauteng, South Africa We are looking for an iOS developer responsible for the development and maintenance of applications aimed at a range of iOS devices including mobile phones and tablet… #jobs #recruiting #careers	Y	Y	Y	
6	Charles Taylor PLC are now recruiting Senior #Analyst .NET #Developers to join their business to build new innovative systems for insurance industry and support existing #IT applications. Location London. Competitive salary + package. Apply #jobs	Y	Y	Y	Y

Table 8: Job-related Postings from Twitter

However, it is also important to note that the nature of social media data is unstructured with respect to conventional job portals. Data from existing commercial job portals are mostly semi-structured in nature which is easier for parsing in JKB but social media data is unstructured in nature and difficult for parsing and entity recognition. Therefore, at the time of crawling, data cleaning will be a challenging task. Next, converting unstructured social media data and presenting in a structured fashion for developing the JKB will be another challenging task. Figure 5 reports a conceptual model to extract relevant and actionable information from unstructured social media data, and subsequently how this knowledge can be disseminated in a concise and structured fashion to jobseekers.

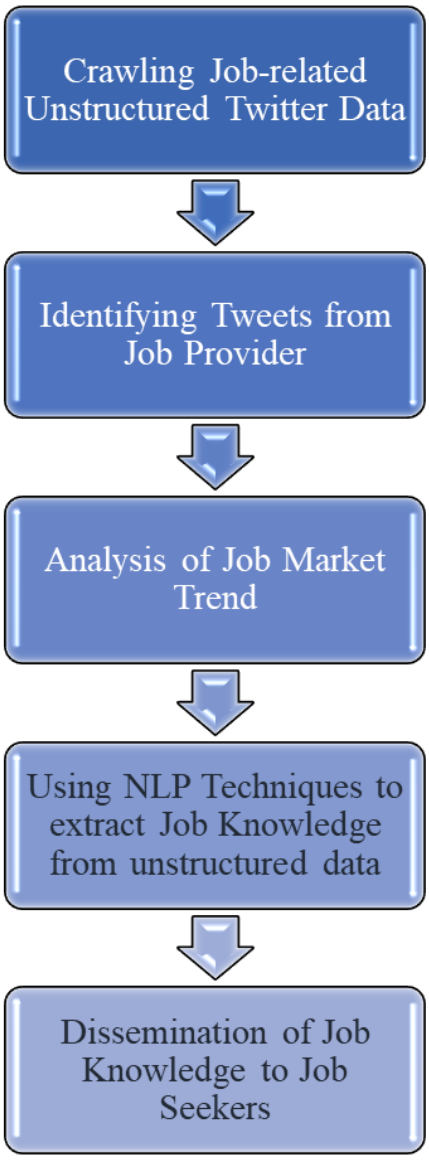

Figure 5: Proposed Approach for extracting Social Media Data for developing JKB

8. Comparison of Proposed Schema of IO1 with ISCO classification

The IO1 report has considered the European Skills, Competences, Qualifications, and Occupations[4] (ESCO), a regulatory body, which identifies and classifies professional occupations, skills, and qualifications relevant for the EU labor market. The proposed schema of IO1 has considered the three most important pillars of ESCO for developing an efficient Job Knowledge Base (JKB). These three pillars are *occupations*, knowledge, *skills* and competences, and *qualifications.* IO1 has noted that the ESCO occupations pillar contains around 3,000 occupation concepts. We observe that each of these occupations is mapped to the ISCO-08 (ISCO: International Standard Classification of Occupations).

The ISCO[5] is "a four-level classification of occupation groups managed by the International Labour Organisation (ILO). Its structure follows a grouping by education level. The two latest versions of ISCO are ISCO-88[6] (dating from 1988) and ISCO-08[7] (dating from 2008)." It is important to note that each occupation identified by ESCO is mapped to exactly one ISCO-08 code. ISCO-08 provides the top four levels for the occupations pillar, and ESCO occupations are located at level 5 and lower. Figure 6 explains 'the role of ISCO 08 in the hierarchical structure of the ESCO occupations pillar'.

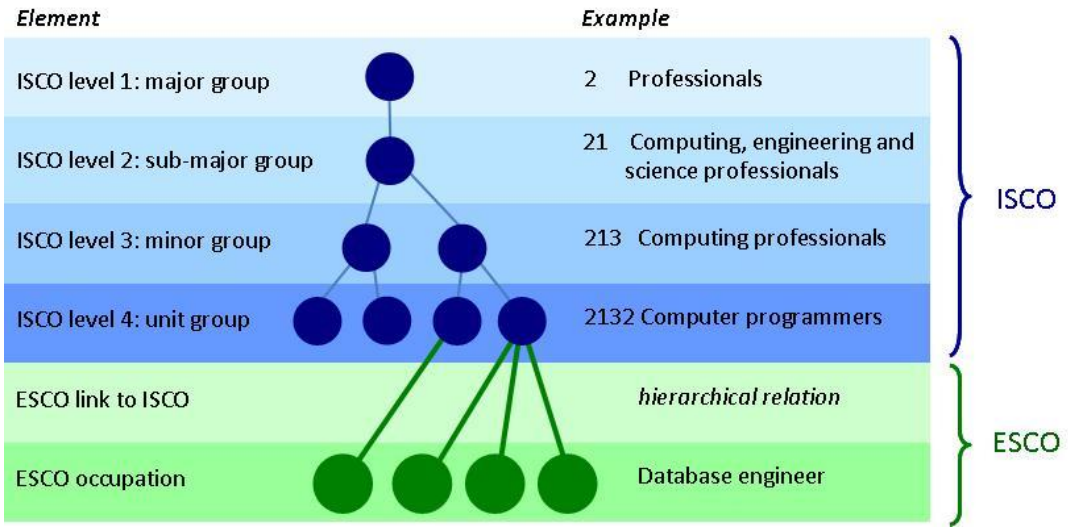

Figure 6: Mapping between ISCO and ESCO *(Source: ESCO website)*

This section of the IO2 report has tried to evaluate this existing ISCO classification in comparison to proposed "*a job category attribute* that can be useful to classify jobs" in IO1. First, we have briefly described the logical flow adopted by ISCO classification for Computer Science related jobs. Next, we have tried to map the ISCO classification as 'a job category attribute' in the context of computer science and data science related jobs.

[4] https://ec.europa.eu/esco/portal/home
[5] https://www.ilo.org/public/english/bureau/stat/isco/
[6] https://www.ilo.org/public/english/bureau/stat/isco/isco88/index.htm
[7] https://www.ilo.org/public/english/bureau/stat/isco/isco08/

In the initial level, ISCO identifies ten occupations such as Managers, **Professionals (Code 2),** Technicians and Associate Professionals (Code 3), Service and sales workers and so on. Computer Science jobs mostly come under the 'Professionals' category and a few jobs also come under the 'Technicians and Associate Professionals' category. According to ISCO, "Professionals increase the existing stock of knowledge" by applying "scientific or artistic concepts and theories". Professionals conduct analysis and research, develop concepts, theories and operational methods. Professionals apply existing knowledge base in the domain of physical sciences, mathematics, engineering and technology etc. Subsequently, occupations in this major 'Professional Group' are classified into sub-major groups such as Science and Engineering Professionals, Health Professionals, **Information and Communications Technology Professionals (Code 25)**, and so on.

According to ISCO classification, "Information and communications technology professionals conduct research; plan, design, write, test, provide advice and improve information technology systems, hardware, software and related concepts for specific applications" in addition to other works. Next, occupations in this sub-major group (i.e., Code 25) are classified into two minor groups, and subsequently into the narrower ISCO groups. The overall hierarchical structure can be represented as follows:

- **Code 2: Professionals**
    - **Code 25: Information and Communications Technology Professionals**
        - **Code 251: Software and Applications Developers and Analysts**
            - Code 2511 Systems Analysts
            - Code 2512 Software Developers
            - Code 2513 Web and Multimedia Developers
            - Code 2514 Applications Programmers
        - **Code 252 Database and Network Professionals**
            - Code 2521 Database Designers and Administrators
            - Code 2522 Systems Administrators
            - Code 2523 Computer Network Professionals
            - Code 2929 Database and network professionals not elsewhere classified

However, as we noted earlier, some of the computer science-related jobs also come under the category **Technicians and Associate Professionals (Code 3).** According to ISCO, Code 3 is mostly about performing "technical and related tasks connected with research and the application of scientific or artistic concepts and operational methods". So, tasks "performed by technicians and associate professionals usually include undertaking and carrying out technical work connected with research and the application of concepts and operational methods in the fields of physical sciences including engineering and technology" among others. So, the overall ISCO hierarchical structure for this category is as follows:

- **Code 3: Technicians and Associate Professionals**
  - **Code 35 Information and Communications Technicians**
    - **Code 351 Information and Communications Technology Operations and User Support Technicians**
      - Code 3511 Information and Communications Technology Operations Technicians
      - Code 3512 Information and Communications Technology User Support Technicians
      - Code 3513 Computer Network and Systems Technicians
      - Code 3514 Web Technicians
    - **Code 352 Telecommunications and Broadcasting Technicians**
      - Code 3521 Broadcasting and Audiovisual Technicians
      - Code 3522 Telecommunications Engineering Technicians

Table 9 reports the possible mapping of ISCO-08 classification with computer science jobs. As we have noted earlier, that some of the job postings are using generic phrases such as, computer science or information technology whereas other job postings are using specific phrases searches data sense or web development. We have tried to map ISCO-08 classifications with job postings from various portals. Our analysis suggests that understanding the specific nature of jobs from generic phrases or upper level categorisation (such as Code 25) is problematic. More importantly, mapping of ISCO-08 classification with specific phrases such as 'data science' also fails to convey the exact nature of the job. For instance, data science jobs can have multiple requirements, but none of the existing classifications of ISCO-08 can capture them properly. In short, first 4 levels of ISCO fails to convey the exact nature of jobs, and we have to go for subsequent layers of ESCO classification. Hence, the lack of proper job category classification in ISCO-08 justifies the proposed "job category attribute that can be useful to classify jobs" in our IO1 report.

Sl. No.	Domain	ISCO-08 Code
1	Computer Science	**Code 25** Information and Communications Technology Professionals
2	Information Technology	**Code 25** Information and Communications Technology Professionals
3	Data Science	**Code 252** Database and Network Professionals **Code 2521** Database Designers and Administrators **Code 2929** Database and network professionals not elsewhere classified
4	Web Development	**Code 2513** Web and Multimedia Developers **Code 3514** Web Technicians

Table 9: Mapping of ISCO 8 Classification with Computer Science Jobs

9. Comparison of Proposed Schema with Existing Job Portals

According to our DISKOW project proposal, we have to compare different available job posting-related standards like HR-XMl, IEEE 1484.20.1, and Schema.org. The objective of studying various standards is to understand the most suitable and comprehensive data model for DISKOW project. IO1 has already explored these standards and tried to understand the advantages and disadvantages of these standards. Finally, IO1 has broadly considered the schema.org model and added a few additional attributes from HR-XML as the final data model. These two additional attributes, from HR-XML, are *'job category'* which will be useful to classify jobs, and *'job positions'* attribute to describe the number of positions available with respect to a given job posting.

Hence, this section of the IO2 report is attempting to compare the schema and existing job portals. To do this analysis, we have considered the Schema.org as our benchmark. Leading tech giants, such as Google, Microsoft, Pinterest, Yandex, and others, came together to develop this Schema.org. It is interesting to note that over 10 million sites use Schema.org to markup their web content and other web-based products like email. In addition to Schema.org, we have also considered the two proposed attributes of IO1.

Table 10 is comparing some commercial job portals with schema.org. Schema.org has identified a long list of attributes in the context of the job posting. However, most of the job portals club them under the broad heading of 'job description.' More importantly, this 'job description' field is very ambiguous and not structured properly. It mostly depends on the recruiter. So, lots of relevant information get merged into this 'job description' field. This is also true for two European Job portals - EURAXESS and EURES. According to our analysis, we feel that all these job portals share most of the Schema.org attributes under the broad heading 'job description'; however, most job postings do not mention the salary/wages. We found most job postings (in existing job portals as well as social media platform) don't report the 'job category' and 'job postings' - as identified in IO1 report.

Our overall analysis, in this IO2 report, suggests that job classification attribute helps us to classify a job more efficiently. So, in the previous section of this IO2 report, we have tried to map job postings with existing ISCO-08 classification. This mapping between job posting and the ISCO-08 job classification was evaluated in the computer science domain, and we noted that ISCO-08 classification fails to correctly classify the jobs such as Data Science jobs. In conclusion, we can conclude that the extended model of IO1 is a very comprehensive data model, and it will be able to capture all details for a particular job posting.

Hence, the next module of this project, i.e., Output 3 (O3), is aiming to develop an advanced technique to extract the knowledge/information/attributes from all the generic field 'job description,' and next, it will try to map this information with the structured and well-defined attributes of Schema.org. This will address the present shortcoming of the generic field known as 'job description'.

Sl. No.	Job-related attributes from Schema.org	Profesia	JobsinMilan	Stepstone.de	Facebook	Corresponding Attributes From	
						EURAXESS	EURES
1	applicantLocationRequirements						
2	applicationContact	Y	Y	Y		Reference Number	Job Id
3	baseSalary	Y			Y		
4	datePosted	Y	Y	Y			Published On
5	educationRequirements	Y	Y	Y		Skills/Qualifications	Job Description
6	employerOverview	Y	Y				Employer/ Job description
7	employmentType	Y	Y	Y	Y	Type of Contract	Type of Contract/Position
8	employmentUnit						
9	estimatedSalary	Y			Y		
10	experienceRequirements		Y		Y	Researcher Profile	Your Qualifications
11	hiringOrganization	Y	Y	Y		Organisation/Company	Employer
12	incentiveCompensation						
13	industry	Y	Y			EU Research Framework Programme	Industry
14	jobBenefits			Y	Y		Job Description
15	jobImmediateStart					Offer Starting Date	
16	jobLocation	Y	Y	Y	Y	Location	workplace
17	jobLocationType						workplace
18	jobStartDate	Y				Offer Starting Date	
19	occupationalCategory					Research Field	
20	physicalRequirement						
21	qualifications	Y	Y	Y		Qualifications	Your Qualifications
22	relevantOccupation						
23	responsibilities		Y	Y			Category/ Your Tasks
24	salaryCurrency	Y					
25	securityClearanceRequirement						
26	sensoryRequirement						
27	skills	Y	Y	Y	Y	Skills/Qualifications	Job description
28	specialCommitments						Job description
29	title	Y	Y	Y	Y	Job Title	Job Title
30	totalJobOpenings						
31	validThrough					Application Deadline	
32	workHours					Hours Per Week	Hours
33	**job category**						
34	**job positions**						

Table 10: Comparison of Job Portals with Schema.Org

Source: Above data-points were extracted during Feb 2020 from their respective websites

10. Conclusion

This report has presented the outcome and the progress accomplished within the scope of the DISKOW IO2. This intellectual output (IO2) report has made an attempt to address two major assigned tasks according to the Description of Work in the DISKOW Grant Agreement. The first task was the identification and classification of JK Data Sources from the web. Hence, this report has rigorously explored different data sources, especially commercial job portals, from the web. As mentioned in the agreement, this report has given more emphasis to the data sources, which are potentially more useful to develop the JKB. This report has also classified available data sources into Global, European, and Local data sources. Accordingly, this report has identified job posting data sources at the Global level, EU level, and Country level. In addition, this report has also explored open data sources such as social media. This report has compared semi-structured data from job portals and unstructured data from social media in the context of job-related postings. This report has also identified some of the major shortcomings of conventional job portals. One of the project partners has successfully coordinated and made an agreement with a country-specific job portal for accessing their database to develop and evaluate our algorithms in the subsequent module within the scope of the DISKOW project.

The second task of the IO2 was the evaluation and verification of open standards for JKB implementation. In continuation of the IO1 report, this report has tried to evaluate our data sources and compared them with open standards such as HR-XML and Data Model for Reusable Competency Definitions (IEEE 1484.20.1). Hence, the second part of this report has evaluated the proposed job posting schema of IO1 in the context of dominant job portals. This report has also explored social media data and tried to understand which all attributes (with respect to the proposed schema) can be extracted from social media data. Based on the analysis, this report has concluded that datasets of existing job portals can be mapped with the proposed data model (of IO1) for developing the JKB. Finally, this report expects that this proposed data model will improve the narrow search capability, and consequently, it will help the subsequent modules to develop an efficient JKB.