Exploring potential of online CV data for labour market intelligence

The case of Europass

Contents

POLICY BACKGROUND **1**

EVIDENCE **2**

1. Data sources 2
2. Biases and assumptions 4
3. Occupations analysis 9
4. Skills 12

CONCLUSIONS **16**

1. Key findings 16
2. Focus on strong points of the data set 16
3. Work experience differs across occupations 16
4. Europass CV data relates well to other sources 17
5. Europass a valid data source for ESCO update 17
6. Mapping unstructured text to frameworks and classifications helps labour market research 18
7. Recommendations on using online CV data for future research 18
8. Developing more open-source software to assist labour market research 19

REFERENCES **19**

Policy background

Europass was launched in 2004 by the European Commission to increase transparency of qualifications and promote mobility of European citizens; it was established by Decision No 2241/2004/EC. In 2018, a new Decision (EU 2018/646) established a common framework for the provision of better services for skills and qualifications (Europass).

From 2005 to 2020, Cedefop designed, developed and maintained the Europass web resources. An online CV editor, progressively available in 28 languages, was developed. About 165 million CVs were generated through the editor during this period.

Cedefop maintained a database of statistical data on the CVs being created, anonymised and processed in accordance with the provisions of Regulation (EU) 2018/1725 on the protection of personal data. The data fields kept included:

- country of residence, age group, and nationality.
- education and training (name of qualification(s), name of education and training institution(s), dates of studies),
- work experience (job title, dates of employment),
- job-related and transversal skills, languages (name of language and CEFR level),

These data offer an opportunity for a wealth of insights and intelligence to be extracted, for example:

a) How do skills change as work experience increases?
b) What are the common skills per occupation?
c) How do common career paths develop?

Cedefop explored the statistics related to labour force characteristics using the data set of 10 million CVs. The analysis aimed to:

a) investigate to what extent the type and quality of information encoded in Europass CVs can provide or indicate a plausible, granular picture of the labour force or parts of it.
b) expose the data set's biases and limitations;

These data include information on a variety of labour force aspects – such as region, occupation, start and end of employment, most recent jobs held – that labour force surveys do not typically gather and cannot easily obtain. Algorithms were constructed to map the free form text of job titles and skills into standard classification codes of the ISCO/ESCO taxonomy.

The necessary data cleansing, processing, and classification were performed using a data pipeline based on the R statistical computing language. The analysis is presented in reproductible form whereby interlinked raw data, analysis, code and results can be inspected and independently reproduced by researchers.

The research was carried out by the Eworx company (Service Contract No 2021-0081-NP-DSI-PHT-ASIA-CVDataAnalytics&Intelligence-008-20), which supported Cedefop in the preparation of this publication. The detailed results, the full text of the analysis (referenced as "Final report") and a number of interactive dashboards and tools are available at https://pub5600.cedefop.europa.eu.

Labour market supply vs demand - Skills intelligence? -> Jasper?

Evidence

1. Data sources

1.1 Europass online CVs

The Europass CV editor online application allows users to create, store and share their curriculum vitae. Cedefop has developed and maintained the Europass application from 2005 to 2020; during this period around 165 million CVs were created.

CVs created by visitors to the Europass CV editor application serve as the main source of data. The Europass statistical database used in the analysis contains anonymised data of more than 10 million CVs created through the Europass CV editor between 2017 and the second quarter of 2020.

Europass CVs contain demographic data, such as sex, country of residence, work experience, education and training, languages and skills. In conformity with EU data protection provisions, information that can identify users, such as free-text entries about skills, has not been preserved.

Cedefop also conducted a project between June and September 2019, where almost 400 000 Europass users offered full access to their CV, on a voluntary basis, including the full text of their skills and descriptions of their work experience. All personal data were aggregated, anonymised and erased 6 months after the completion of the analysis.

1.2 European skills, competences, qualifications and occupations (ESCO)

ESCO is the European multilingual classification of skills, competences, qualifications and occupations.

ESCO documents relationships between the different concepts, such as the skills required for a job or occupation, and establishes a hierarchy for the defined occupation concepts based on ISCO.

The ESCO/ISCO version 1.0.8, used in this analysis, contains 2 942 occupations and 13 685 skills/competences and knowledge concepts, included in the ESCO taxonomy, with labels and descriptions translated into 27 languages.

1.3 European qualifications framework (EQF)

EQF is the European reference tool for the description and comparison of qualifications developed at national, international or sectoral level.

It consists of eight levels, expressed as learning outcomes (knowledge, skills and responsibility and autonomy) with increasing levels of proficiency.

A machine learning classifier was used to align qualifications contained in the Europass CV with EQF levels.

1.4 International standard classification of education (ISCED)

ISCED is a framework adopted by UNESCO in 2011 to classify educational activities, as defined in programmes, and the resulting qualifications into internationally agreed categories (levels and fields of education).

The ISCED fields of education and training classification (ISCED-F2013) was used in this study.

1.5 European Union labour force survey (EU-LFS)

EU-LFS is a large-scale household sample survey of the labour participation of people aged 15 and over. It is conducted across all Member States of the European Union, as well as four candidate countries (Montenegro, Republic of North Macedonia, Serbia and Türkiye) and three EFTA countries (Iceland, Norway and Switzerland).

Specific indicators from EU-LFS have been selected for comparisons with equivalent measurements in the Europass CV data:

(a) employment by sex, age, professional status and occupation;

(b) previous occupations of the unemployed, by sex;

(c) long-term unemployment (at least one year) by sex, age, educational attainment level and NUTS 2 regions;

(d) employment by sex, age and job tenure;

(e) job tenure by sex, age, professional status and occupation.

1.6 Skills in online job advertisements (Skills-OVATE)

Cedefop's Skills-OVATE offers detailed information on the jobs and skills employers demand based on online job advertisements (OJAs) in 28 European countries.

It is based on the analysis of more than 100 million online job ads between July 2018 and December 2020.

Skills-OVATE is powered by Cedefop's and Eurostat's joint work in the context of the Web Intelligence Hub.

The Skills-OVATE database was used as a reference for this project to experiment with skills gap analysis as well as for a comparison at concept level.

2. Biases and assumptions

The main flaw of the data subject to this analysis is the bias with respect to **coverage** and **representativeness** of the labour force. CVs generated online are not fully representative of the European labour force: Europass is not equally used among different breakdowns of people searching for employment - either trying to get a new job, unemployed or entering labour market for the first time. In addition, the use of Europass is uneven depending on the country.

These problems of bias and coverage exist for all online sources of labour market data, both for supply and demand; this has also been underlined in the literature on online job vacancies.

2.1 Working with CV data

The Europass CV data are not elicited or generated as part of any research project. The data are specifically generated through user interaction with an online tool, so the quality of data is subject to user behaviour.

Information gathered through the Europass CV editor is likely to have high non-response or incomplete response bias - more so than in traditional surveys. This can be attributed to different factors, including incomplete CVs left for editing at a later stage and never completed. The Europass CV editor may also store multiple versions of a CV, making it difficult to distinguish completed CVs from draft documents. To reduce this 'noise', a deduplication process was performed, and only the latest CV associated with each unique reported email address was considered for the analysis.

As most fields on the Europass CV editor are optional and there is no limitation in the number of characters when providing the description of skills or tasks, the completion rate for specific information may vary by user and by pillar (See Table 1), even after deduplication. Given that most CVs are often created to apply for a specific job, a user may choose to not disclose certain information (e.g. age and sex) to a potential employer and may not consider every field as relevant for their application. Certain fields may also require familiarity with standards (e.g. EQF) that many users may not be aware of and may leave blank or fill erroneously. Absence of data is common. Some

of the missing data (e.g. EQF level and ESCO occupations) in the Europass data set were imputed in the cleansing process based on related information shared for other fields.

One more source of bias that is inherent to studies that rely on self-reporting data, and is especially prevalent from a source of self-created CVs, is response or recall bias. The accuracy and completeness of users' recollection of their past work experience cannot be guaranteed, and details are expected to be omitted, especially for CVs with extensive work experience.

Finally, people with diverse work experience may significantly tailor their CV when applying for different jobs. Past work experience not relevant to a user's current career trajectory may not be included and the CV may then not capture the full range of an individual's experiences, skills or qualifications.

Table 1. Completion Rate of CV Fields by Pillar

Pillar	CV Field	Completion Rate
Demographics	CV Language	100%
Demographics	Creation Date	100%
Demographics	Country	96%
Demographics	Birth Year	50%
Demographics	Gender	44%
Work Experiences	Recruitment Year (Start of employment)	97%
Work Experiences	Termination Year (End of employment)	95%
Work Experiences	Job title (Label)	75%
Work Experiences	Job title (ESCO classification)	22%
Work Experiences	Employer	8%
Qualifications	Qualification title (Label)	91%
Qualifications	Enrolment Year (Start of studies/training)	90%
Qualifications	Organisation Country	79%
Qualifications	Graduation Year	79%
Qualifications	EQF Level	12%
Qualifications	Educational Field (ISCED-F)	3%

2.2 Technological diffusion

Given the period of data collection (2017 to 2020), a source of time-varying bias on the data set is related to digitalisation. The switch from 'offline' resources (e.g. offline text editors) to free and/or low-cost online tools (Europass, LinkedIn, Indeed, etc.) is linked to the wide use of the internet for job vacancies (demand) and applications (supply).

On supply, Barnichon (2010) demonstrates that this shift in vacancy coverage closely resembles the S-shape typical of technology diffusion in the United States, as well as the similarly S-shaped fraction of internet users in that country. Similar results are expected for online CVs, leading to increasing coverage in time. For example, in 2020 cooks may use online CV editors more frequently in 2017, showing a false relative increase of cooks within this time span. To some extent, this effect can be smoothed out by comparing ratios, relative changes, or by introducing time-dependent reweighting.

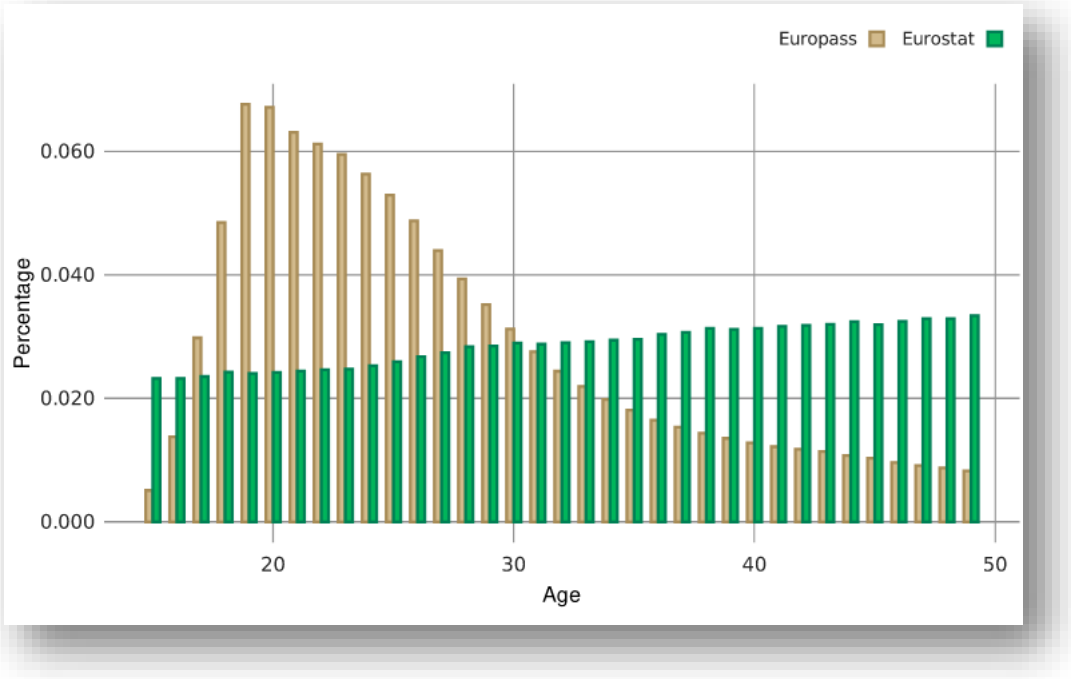
2.3 Europass users compared to the labour force

The extent to which Europass user characteristics are biased in relation to the actual labour force is hard to quantify, even for well represented categories in the data set. The creation of a CV is most often linked to an individual's intention to apply for a job. Businesses may use tailored templates, developed for their specific needs, in which case a CV created on Europass will not be of use. Also, some countries may promote Europass more actively than others, resulting in a significant discrepancy in the level of use of the tool in each country.

Medium- and high-skilled citizens actively looking for a job use the Europass CV editor more than other segments, as the corresponding job vacancies demand higher quality CVs, motivating candidates to use online editors that better meet employer requirements.

Internet resources such as the Europass CV editor are used by younger, educated generations, who are more familiar with digital tools and are naturally more likely to be jobseekers. Those two biases are entangled and lead to an overrepresentation of younger people in the data set. Specifically, we find a mean age of 29, with 72% of users under 32 years old. Restricting the analysis to a narrow range of ages will likely elicit better results.

Figure 1. **Comparing Europass user age distribution to the Eurostat demographics indicator**



Source: Cedefop. Europass statistical database 2017-20.

Figure 2. **Distribution of country of residence among Europass users**

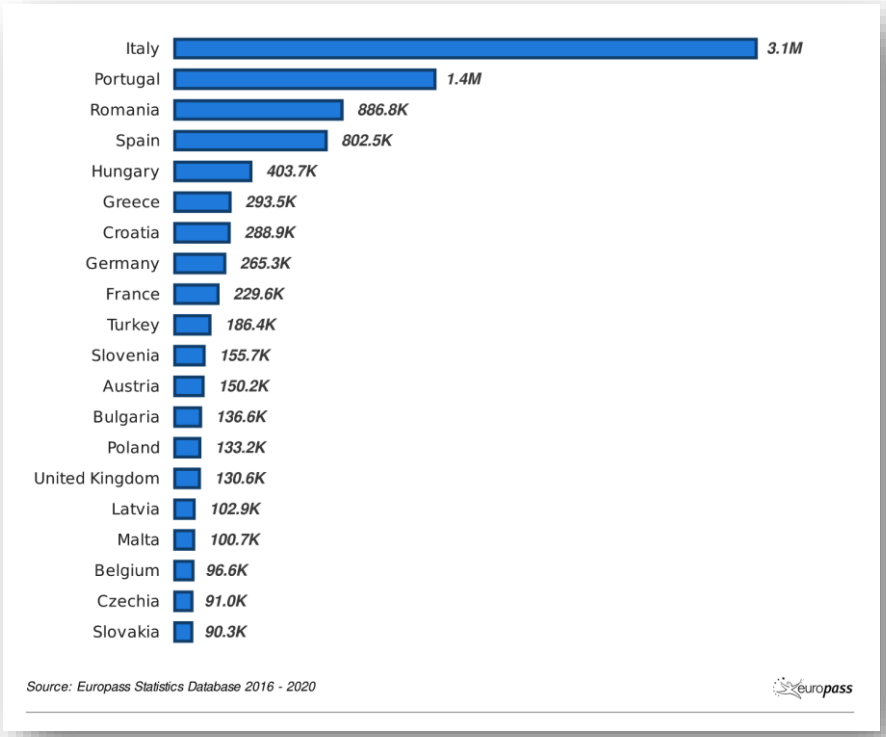
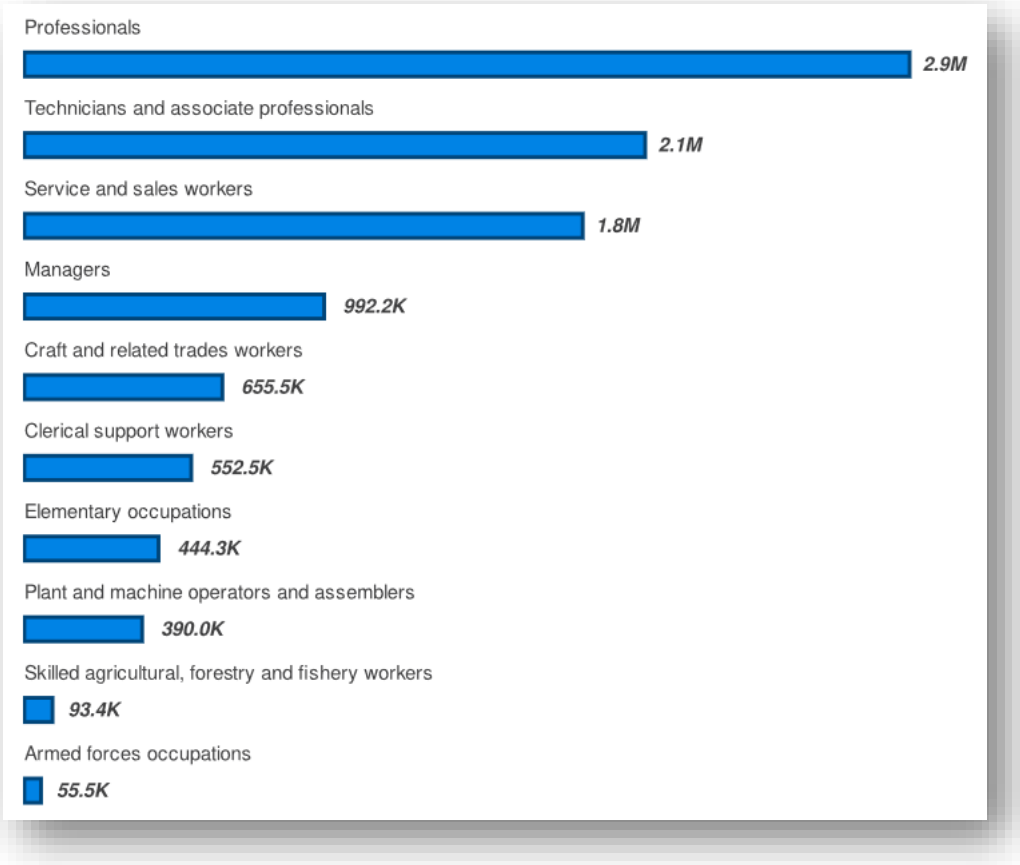


Figure 3. Distribution of occupations of Europass users (ISCO level one)



Source: Cedefop. Europass statistical database 2017-20.

Though the Europass CV editor application is available internationally, it is used more in European Union countries; even here, it is not evenly adopted across all countries. Italy and Portugal are the two countries with the highest percentage of the population using the Europass CV editor: Portugal government has been actively promoting the Europass CV since its launch, resulting in an early and massive adoption by the population; in Italy, Europass is administrated at regional level, allowing targeted, proximity promotion campaigns. On the contrary, countries like Denmark and Poland have a particularly small sample in proportion to their population, partly due to a lower rate of employment, leading to a low emigration rate. Deriving statistics for countries with a larger sample size is generally more feasible with this type of data set. Weighting can also be used to produce a general European indicator. We derive a restricted weighting scheme based on countries in EA-19.

2.4 Biases in calculation of the changing of Skills requirements

One way to measure how the skill requirements vary depending on the job is to compare skills reported by younger users with those of older users with the same job position. There are some limitations to consider when applying this approach:

1. The data is based on self-reports, which may not accurately reflect the actual skill requirements of the job.

2. Younger and older workers may have different approaches to reporting skills, which could introduce bias.

3. The industry itself might be undergoing transformations that affect skill requirements, making it difficult to attribute changes solely to generational differences.

4. If Europass is more popular among certain age groups as it is the case here, the data may not be representative for older groups.

5. Older workers reporting current skills might have adapted to new technologies and methods, thus their skill set might not accurately reflect the requirements when they first entered the role.

6. Other variables like education, geographic location, and company size might also play a role in the skills that are reported.

7. The definition or understanding of what constitutes a "skill" might differ across age groups.

8. Cross-sectional vs Longitudinal Data: a more robust approach might be to use longitudinal data to track how the skills required for the same job change over time.

3. Occupations analysis

The analysis of job titles shows that the yearly distribution of recruitments per ISCO group in the European countries (EA-19) changed in the period under review (2000 to 2019).

The representation of some occupations (e.g. ISCO 2 groups: food preparation assistants, personal care workers, health professionals, and personal service workers) increased in CVs (Figure 6).

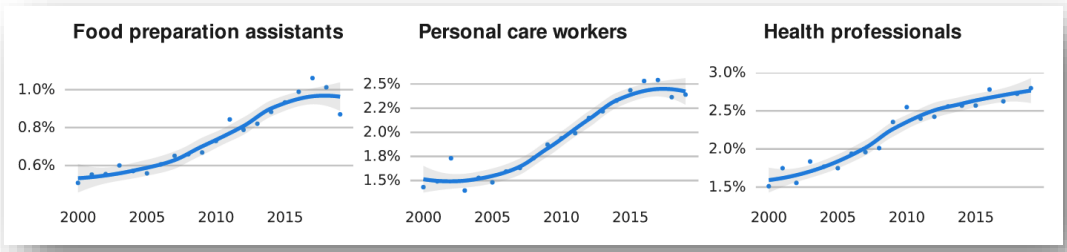
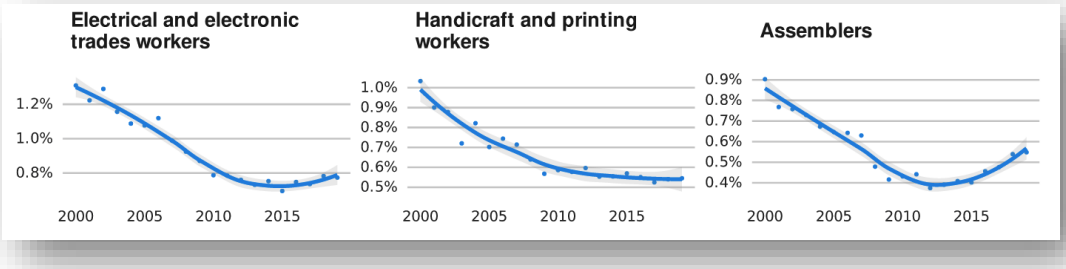


Figure 4. **Example of ISCO 2 occupations increasing in distribution of recruitments for EA-19**

Source: Cedefop. Europass statistical database 2017-20.

In contrast, the representation of other occupations decreased (e.g ISCO 2 groups: electrical and electronic trades workers, assemblers, building and related trades workers, and general and keyboard clerks).

Figure 5. **Example of ISCO 2 occupations decreasing in distribution of recruitments for EA-19**



Source: Cedefop. Europass statistical database 2017-20.

Recruitment trends between 2017 and 2020, as observed by the EU labour force survey, are confirmed for occupations that are well-represented in the Europass data set.

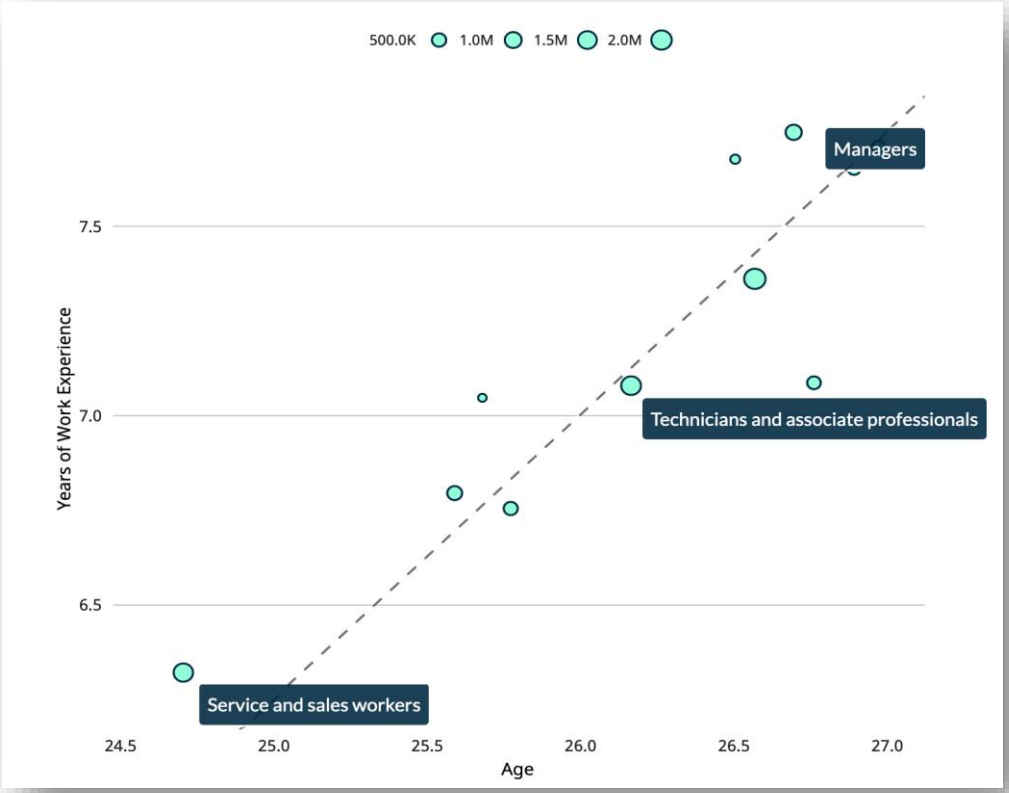
The odds ratio change between the two sources aligns within less than 1% deviation for managers, professionals, plant and machine operators and assemblers, and service and sales workers. For most other occupations, the trend's magnitude is captured with a larger deviation, but its direction still aligns on most cases.

3.1 Relationship between work experience and age (by type of occupation)

The analysis of career paths described on Europass CVs shows that the volume of work experience varies from occupation to occupation. This finding suggests a significant variation in the career stability of different groups of users.

Europass provides not only the current employment status of its users but also a window to their past labour market participation, helping to identify the more disadvantaged occupation groups and unravelling their demographic characteristics.

Figure 6. **Relationship between age and years of work experience with ISCO 1 occupations placed based on their mean values**



Source: Cedefop. Europass statistical database 2017-20.

Each green circle in the above figure represents an ISCO 1 occupation. The occupations placed to the left in the x-axis are for ISCO 1 occupations is more commonly seen among younger individuals (e.g. service and sales workers), while those to the right are more commonly professed by relatively older individuals (e.g. managers). The occupations that are upmost on the y-axis are for ISCO 1 occupations whose users had relatively more years of work experience at the time of recruitment (e.g. managers), whereas those that lie on the bottom are entry-level jobs (e.g. service and sales workers).

Looking at different ISCO hierarchical levels, the relationship between age and work experience is strongly linear. When we look at recruitment type as a relationship between age and work experience, we see that young people with less work experience are more likely to get hired as waiters or serve in the army. As we move across the fitted line to the right, we gradually encounter vocations that require more work experience such as managers. Above the line are more likely skill-based occupations that allow entry into the labour market at a younger age and/or occupations that require more experience to be recruited like managers. Those below the line are more likely to be knowledge-based and/or require less experience for recruitment, such as school teachers, life science experts, and programmers.

The deviation between the dotted line and each occupation is a measure of how the amount of work experience an individual recruited for an ISCO 1 occupation has, compared to the expected level of experience. For example, managers appear above the line, which means that an individual starting work as a manager has more years of work experience than the average newly recruited individual.

We see an opposite trend for elementary occupations, whose users have fewer years of work experience than expected.

4. Skills

Along with work experience, the users of Europass describe their skills in free-text paragraphs. The text of these sections is matched to a classification in the ESCO model through our data cleansing process. Apart from matching with ESCO, specific keywords in the free text are also extracted and assessed.

4.1 Distribution of skills

The distribution of non-linguistic skills observed overall is reflected in the following graph, with the number shown representing the total number of CV skill entries matched for each ESCO skill group.

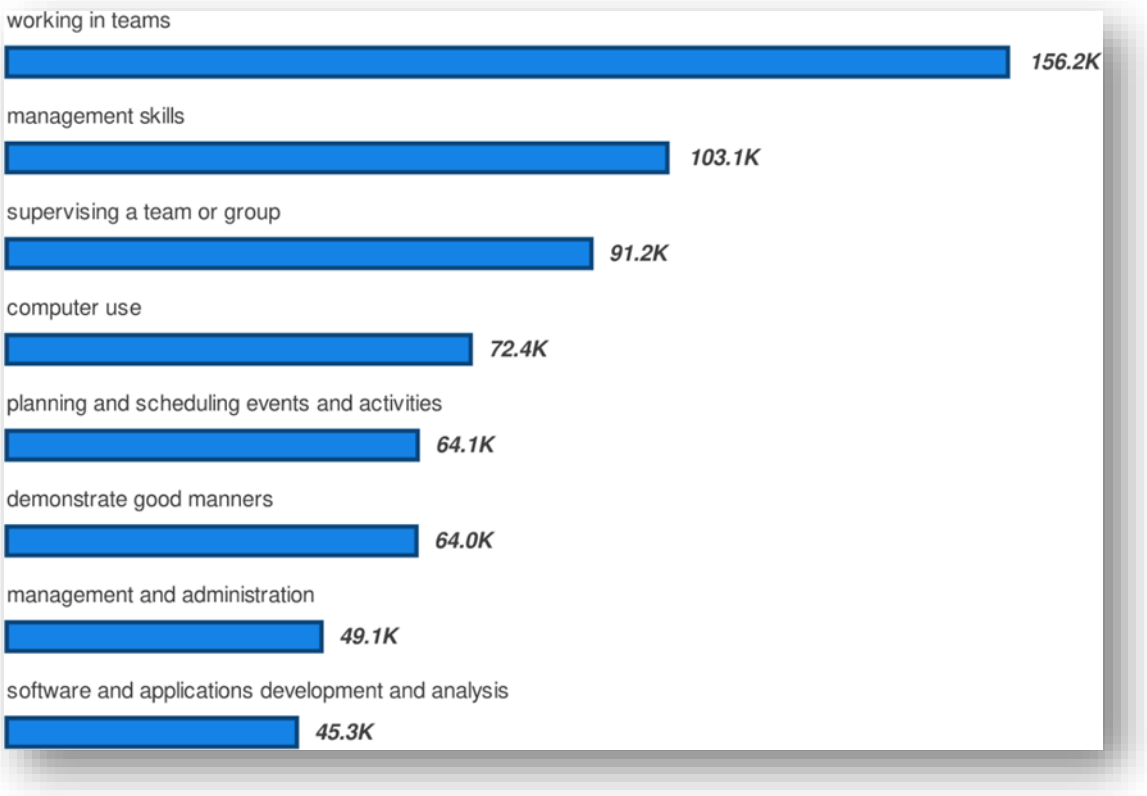
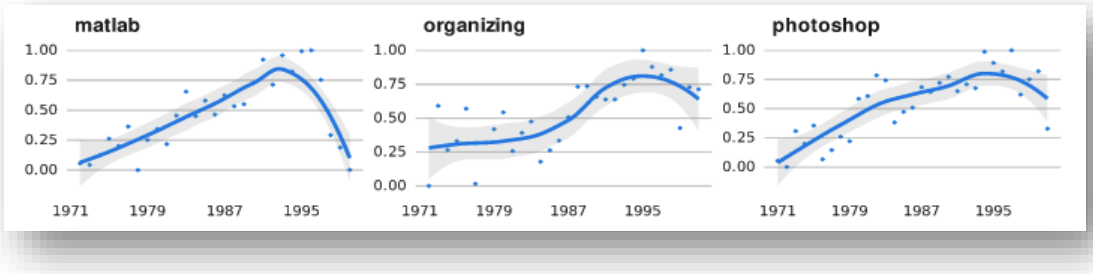


Figure XX. Distribution of ESCO skills found on Europass CVs

4.2 Skills by occupation and birth year

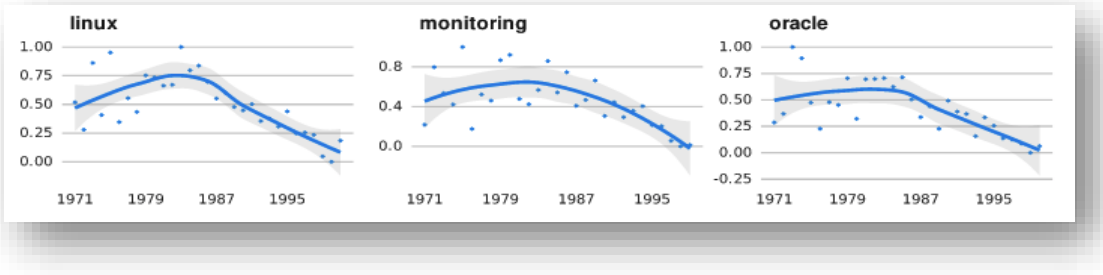
One way to measure how the skill requirements of each job are changing, is to compare skills reported by younger users with skills of users with more experience who report the same job position. For the related limitations of this approach, see section **Error! Reference source not found.**. We calculated the birth year (or age) distribution of each skill per ISCO occupation and performed a regression analysis. In this case, the slope of the fitted statistical model is an indicator of the relative increase/decrease of the inclusion of a skill as the year of birth increases.

Figure 7. **Example of keywords more often included in younger users' skills with latest job, professionals**



Source: Cedefop. Europass statistical database 2017-20.

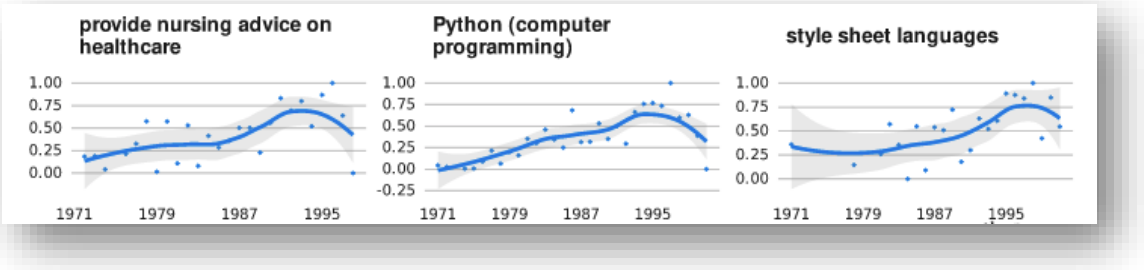
Figure 8**.** **Example of keywords more often included in older users' skills for their latest job, professionals**



Source: Cedefop. Europass statistical database 2017-20.

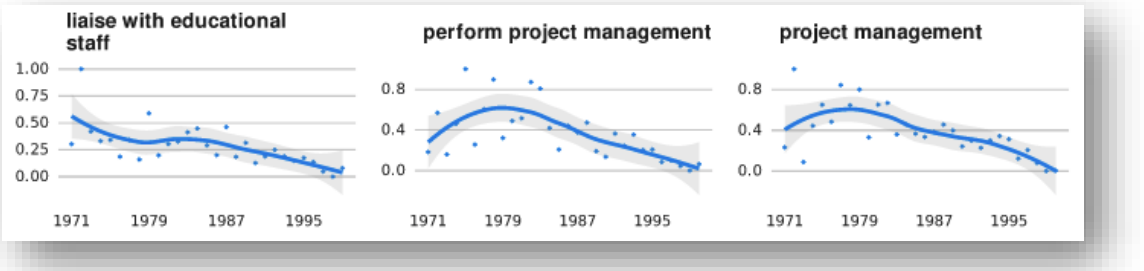
Keywords like Photoshop™, Python™ and MATLAB™ display a positive trend for ISCO 1 group 'professionals' with respect to birth year, meaning that they are more commonly reported by younger users compared to older ones, who mention keywords like Oracle™, application, and security.

Figure 9. **Example of ESCO skills more often entered by younger users for their latest job, professionals**



Source: Cedefop. Europass statistical database 2017-20.

Figure 10. **Example of ESCO skills more often entered by older users for their latest job, professionals**



Source: Cedefop. Europass statistical database 2017-20.

Managerial ESCO skills (e.g. project management, draft corporate emails) are more frequently reported among older users, while ESCO skills related to childcare and students (e.g. assist children with homework, communicate with young people) are more frequently observed among younger people.

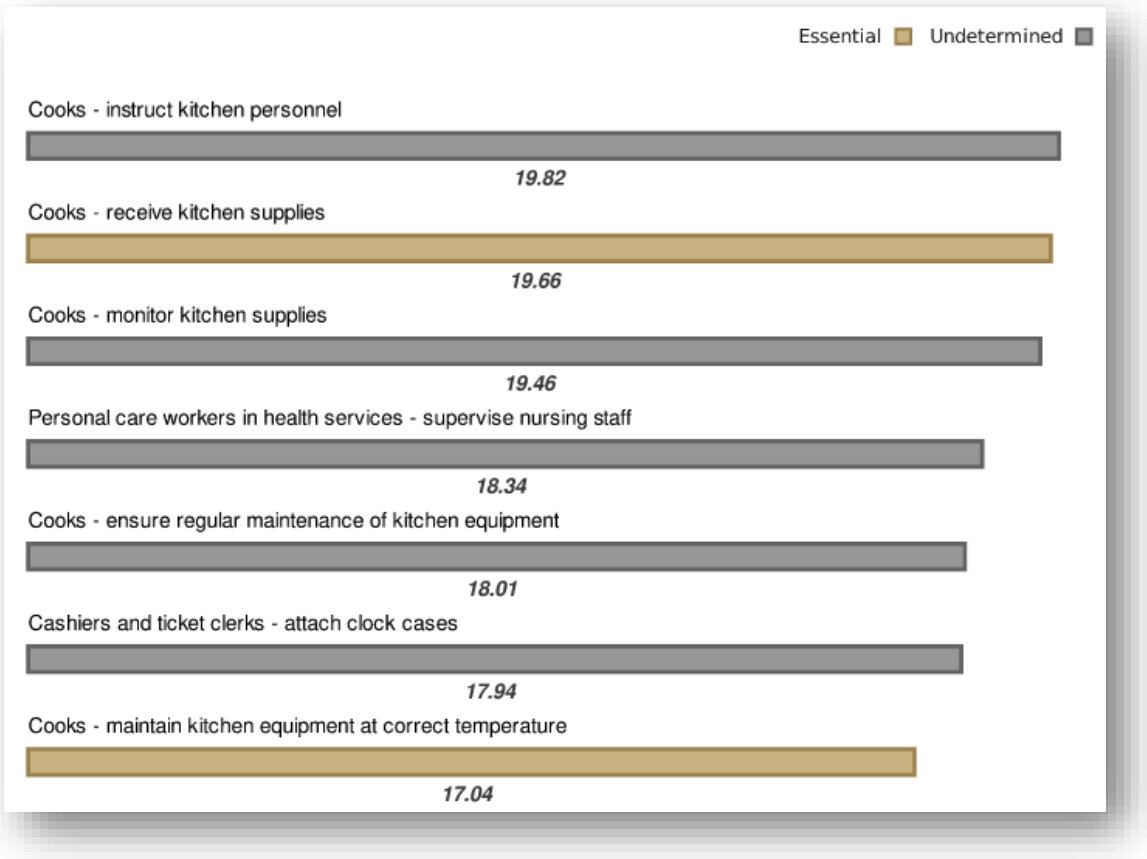
4.3 Associations between skills and occupations

Having for each CV the latest occupation and the related skills, listed in free text, we can assess the "strength" of the relation between the two (skills and occupation). Through association rules mining, the relationship between an occupation and a skill has been quantified on a metric called 'lift' (see Section 5.4.1 and Annex A: Methodological Notes - Association rules) of the "Final report". A high lift suggests a strong relationship between a skill and an occupation.

By ordering occupation and skill pairs by lift, we identify the strongest associations between skills and occupations present in the CVs and whether they are already encoded in the ESCO model.

If a skill is within at least one ESCO occupation as 'essential' it is marked as such; otherwise it is marked as 'undetermined'.

Figure 11. **Associations between occupations and skills identified through market basked analysis for service and sales workers**



Source: Cedefop. Europass statistical database 2017-20.

Measuring the inclusion of skills per occupation reveals new meaningful associations that are not encoded in the ESCO model. Examples of ISCO 3 and ESCO skill pairs that do not seem to be encoded in ESCO include: a) occupation "cooks" - skill "instruct kitchen personal"; b) occupation "nursing and midwifery professionals" - skill "communicate with nursing staff"; and c) occupation "domestic, hotel and office cleaners and helpers" - skill "conduct cleaning tasks".

Analysis data from Europass and other similar sources (e.g. with information on real-life jobs and skills utilised or needed in them) can inform and assist researchers in improving the ESCO model by documenting the most common associations.

Conclusions

1. Key findings

The analysis of the CV data yielded the following key findings:

a) Europass CV data provide significant labour market intelligence, especially for highly qualified young jobseekers, the main users of Europass;
b) there is a strong relationship between the type of occupation and career stability and continuity;
c) the analysis of skills contained in Europass CV can help identify skills mismatches, underskilling/underqualification and overskilling/overqualification (e.g. less skills appearing in CVs than those expected for the specific occupation; high-skilled, highly-educated people working on irrelevant or low-skill jobs);
d) there is a clear correlation between knowledge of foreign language skills and transnational mobility;
e) skills contained in the Europass CV reveal additional associations between occupations and job-related skills, which can be used to develop further the ESCO classification;
f) mapping unstructured text contained in Europass CVs to statistical frameworks and classification systems (e.g. occupations) through open-source libraries greatly improves labour market research capabilities;
g) this pilot CV analysis was able to reproduce some of the metrics (employment trends) observed in official statistics (e.g. EU-LFS);
h) there is a strong correlation between the occupations found in Europass CVs and those published in online job vacancies;
i) developing open-source software libraries supports labour market research.

2. Focus on strong points of the data set

The content of Europass CVs under review does not reflect the different groups of the labour force. People under the age of 32 create online CVs much more often than older individuals. Work experiences more typically reported in online CVs tend to require more specialisation, which makes professionals and managers overrepresented. University education level is most frequently observed. Also, the main motivation behind creation of CVs is job search, rather than an attempt to list comprehensively all professional life experiences.

Therefore, generalising the observations made on the Europass dataset to the EU27 labour market proves to be challenging. However, sources like Europass can be used to obtain specific insight into some of the more well-represented subgroups through custom queries.

3. Work experience differs across occupations

The study demonstrates that the relationship between the amount of work experience and the type of occupation can be quantified through the analysis of online CV data.

Using career histories reported on CVs, we find that accumulated work experience varies depending on the occupation (section 5.3.3 of the "Final report").

This finding suggests significant variations in the career stability and ease/difficulty to find a job on the specific field and gain work experience, as well as numerous patterns regarding their entry in the labour market. Analysis of sources like Europass, which provide not only the current employment status of its users but also a window to their past labour market participation, unlocks unique opportunities for identifying the more disadvantaged occupation groups and unravelling their demographic characteristics. Understanding the career paths provides valuable information for policy-makers to mitigate both frictional and structural unemployment, facilitate the integration of specific groups on the labour market and improve career stability. It can also inform vocational planning and associated education and training policies.

4. Europass CV data relates well to other sources

4.1 Comparison with official statistics of EU-LFS

The study was able to reproduce some of the metrics observed in official statistics (EU-LFS). Employment trends over the short period of data collection as calculated using the Europass CV data generally corroborate the findings of other official sources of statistics for population over the age of 25 with respect to ISCO 1 (see sections 6.1 and 6.2 of the "Final report").

Minor year-to-year fluctuations noted on most groups of occupations are observable in Europass data, meaning that online CVs and similar data sources can complement official sources on the identification of such trends and assist survey designers and policy-makers. Measurements over longer timespans using the career histories as they appear on CVs is more challenging due to recall bias.

However, through careful data calibration, focus on indicators of relative change instead of absolute values, and use of custom metrics (e.g. ratios) that help to smooth out some of the bias, patterns may also emerge over longer timespans via exploration of career histories.

4.2 Correlation with online job vacancies

The distribution of occupations observed in online vacancies (Cedefop (2019)) resembles the equivalent distribution observed in the field "Job applied for" (which has alternative labels "Position" and "Preferred job") in Europass CVs (see section 6.3 of the "Final report"). This confirms that the trend of jobs searched for (as stated in CVs) and of available vacancies (as appearing in online advertisements) is similar, as expected, although with differences that are to be closely studied for the degree that they represent a skills gap. It means that making measurements related to the demand side of the labour market online using CV data can not only enhance the already existing sources that measure demand, but also help to identify gaps between supply and demand either at EU Member State level or across the EU as a whole.

Research into this gap and its evolution in the long term can inform policies, for example through new education and training programmes, or adjustments in migration and mobility policies. Validation and cross-referencing with other data sources such as the LFS would be beneficial to enhance the robustness of any inferences made.

5. Europass a valid data source for ESCO update

The identification of skills specific to certain occupations is possible through analysis of Europass CVs. By establishing associations between occupations and skills, applying traditional market basket analysis, it has been possible to confirm the validity of relationships within the ESCO taxonomy, as well as identifying new ones, not yet documented in the ESCO model (see section 5.4 "Skills-to-

Occupations Associations in the ESCO Model Compared to the Collected CV Data" in the "Final report").

Filtering recurrent skills, using indices known from economics (i.e. the revealed comparative advantage (RCA) index), helped to establish which skills are overexpressed (and more important) and for which occupations

6. Mapping unstructured text to frameworks and classifications helps labour market research

Going from unstructured, free text to a restricted number of well-defined classes (e.g. ISCO occupations) is a major challenge when dealing with online CV sources. For this analysis, software was developed that takes user-input free text and matches it to the ESCO classification, as well as to the European qualifications framework and the ISCED Fields of education and training.

This software is published as open-source statistical packages that may help other researchers working with sources similar to Europass. labourR has already been published on CRAN, with educationR and iscoCrosswalks currently available on GitHub and on the early phase of the CRAN submission process.

7. Recommendations on using online CV data for future research

First, the country-level representativeness and reliability of the data source must be assessed, as is the case for job vacancy analysis. Focusing the analysis at country level, a dominant market share of occupations in countries can be viewed as a good source of information that can lead to credible and transferable research findings.

Second, representativeness and reliability of data must be evaluated in relation to a certain study topic. Certain characteristics of coverage and sampling flaws can be addressed using a data segment or sub-sample that can be regarded as representative. These biases are likely to be less pronounced in professions or labour market niches that are heavily exposed to the internet (for example, IT-related professions). Unlike survey data, more detail is provided.

Third, online CV data could be combined with other sources of online vacancy data depending on the research question, since online CV data and other sources of online vacancy data are strongly connected. We note that job sectors are likely to be exposed online similarly from the demand side as well as supply. Online vacancies are also expected to cause a driving effect on online CVs to be ‘tailored’ to the jobs advertised. This leads towards an equilibrium between online job vacancies and online CVs created. Evidence is found in the data set of this study, since the distribution of estimated ISCO codes is highly correlated to that of online job advertisements, despite the well-known gap between labour force supply and demand.

Finally, advanced statistical methods based on missing data, such as model-based approaches to the imputation of data not missing at random, could be employed to remove biases resulting from the structure of online CV data. Also, it would be useful to link this data set to that generated by the new Europass CV editor. We also underline the fact that the Europass CV editor hosted by Cedefop stopped working a couple of months after the COVID pandemic. By comparing these data sets, insights could be derived for the characteristics of the users before and after the pandemic. This is a challenging task since a different UI/UX of the new web application might affect user behaviour and

introduce different sets of biases. However, large effects resulted from the pandemic and the digitisation of the job market might still be observable and quantifiable.

The good news for the arguments on naturally occurring data from online sources is that internet-based applications from both labour market supply and demand (search engines, online CV builders, online job advertisements, and so on) are likely to become the dominant ecosystem for large segments of the labour market used for job matching; this will considerably increase the percentage of workers and firms that participate in it (Askitas and Zimmermann, 2015). In comparison to traditional employment channels and processes, the online job market can offer a greater variety of options and increasingly sophisticated tools for assessing the suitability of a job or a job prospect.

8. Developing more open-source software to assist labour market research

This pilot study highlights the relevance of software libraries for labour market research. Developing, maintaining and supporting open-source software is critical for the future of labour market research. It can help research teams to benefit from the work done by fellow researchers to answer recurring questions that arise for the entire research community in the field; it will also reinforce the presence of labour market researchers in the broader software community for statistical research (e.g. on CRAN), enabling continuous feedback and improvements. Libraries missing from the community include additional text mining/machine learning libraries like labourR and educationR, as well as bindings of relevant APIs (e.g. ESCO) to popular programming languages like R and Python.

References

Agrawal, R.; Imieliński, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, pp. 207-216.

Askitas, N. and Zimmermann, K. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower*, Vol. 36, No 1, pp 2-12.

Barnichon, R. (2010). Building a composite help-wanted index. *Economics Letters,* Vol. 109, No 3.

Boettiger, C. (2015). An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review, Vol. 49, No 1, pp. 71-79.

Boselli, R. et al. (2018). Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, Vol. 86, pp. 319-328.

Buelens, B. et al. (2014). *Selectivity of big data.* Statistics Netherlands Discussion Paper, 201411.

Cedefop (2019). *Online job vacancies and skills analysis: a Cedefop pan-European approach.* Luxembourg: Publications Office.

Chambers, J.M. (1992) Linear models. In: Chambers, J.M. and Hastie, T.J. (eds.). *Statistical Models in S*, Chapter 4. Routledge.

Chan, B.K. (2018). Data analysis using R programming. In: *Biostatistics for Human Genetic Epidemiology*, pp. 47-122. Springer.

Chancellor, S. and Counts, S. (2018). Measuring employment demand using internet search data. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* pp. 1-14.

Chapman, C. and Feit, E.M. (2015). *R for marketing research and analytics.* New York: Springer.

D'Amuri, F. and Marcucci, J. (2010). *'Google It!' Forecasting the US unemployment rate with a Google job search index.* FEEM working paper, No 31.2010.

Fernandez-Sanz, L.; Gomez-Perez, J. and Castillo-Martinez, A. (2017). e-Skills Match. *Computer Standards and Interfaces,* Vol. 51 (C), pp. 30-42.

Guzi, M. and de Pedraza García, P. (2015). A web survey analysis of subjective well-being. *International Journal of Manpower,* Vol. 36 No 1, pp. 48-67.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, Vol. 36, No 3, pp. 119-167.

James, G. et al. (2013). *An introduction to statistical learning.* Springer texts in Statistics. New York: Springer.

Jhaver, S.; Cranshaw, J. and Counts, S. (2019). Measuring professional skill development in US cities using internet search queries. In: *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13, pp. 267-277.

Kennan, M.A. et al. (2006). Changing workplace demands: what job ads tell us. In *Aslib Proceedings,* Vol. 58 No 3, pp. 179-196. Emerald.

Kitchin, Rob. (2015). Big data and official statistics: opportunities, challenges and risks. *SSRN Statistical Journal of the International Association of Official Statistics*, Vol. 31, No 3, pp. 471-481.

Knuth, D. (1984). Literate Programming. California: Stanford University Center for the Study of Language and Information.

Kuhn, P. and Mansour, H. (2014). Is internet job search still ineffective? *The Economic Journal*, Vol. 124, No 581, pp. 1213-1233.

Kwartler, T. (2017). *Text mining in practice with R.* John Wiley.

Muñoz de Bustillo, R. and De Pedraza, P. (2010). Determinants of job insecurity in five European countries. *European Journal of Industrial Relations,* Vol. 16, No 1, pp. 5-20.

Piatetsky and Shapiro, G. (1991). *Discovery, analysis, and presentation of strong rules*. Knowledge discovery in databases, pp. 229-238.

Tukey, J.W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160).

Potter, J. (2002). Two kinds of natural. *Discourse Studies,* Vol. 4, No 4, pp. 539-542.

Pouliakas, K. (2021). *Artificial intelligence and job automation: an EU analysis using online job vacancy data.* Luxembourg: Publications Office. Cedefop working paper, No 6.

Vydra, S. and Klievink, B. (2019). Techno-optimism and policy-pessimism in the public sector big data debate. *Government Information Quarterly*, Vol. 36, No 4, pp. 101383.

Wickham, H. (2014). *Tidy data.* Journal of statistical software, 59(1), 1-23.

Wilkinson, L. (2012). The grammar of graphics. In: *Handbook of computational statistics*, pp. 375-414. Springer.

Wise, S.; Henninger, M. and Kennan, M.A. (2011). Changing trends in LIS job advertisements. *Australian academic and research libraries,* Vol. 42, No 4, pp. 268-295.

Xie, Y. (2015). Dynamic Documents with R and Knitr: 2nd edition. Chapman and Hall/CRC the R Series. Routledge.