ESSnet Big Data

Specific Grant Agreement No 1 (SGA-1)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
http://www.cros-portal.eu/.........

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2015.007-2016.085**

Work Package 1

Web scraping / Job vacancies

Deliverable 1.2

Interim Technical Report (SGA-1)

Version 2016-11-11

Prepared by: Nigel Swier (ONS, UK)

Ingegerd Jansson, Dan Wu (SCB, Sweden)
Boro Nikic (SURS, Slovenia)
Christina Pierrakou (ELSTAT, Greece)
Thomas Körner, Martina Rengers (DESTATIS, Germany)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775

Contents

List of figures

List of tables

1. Introduction

This report summarises the technical progress made with the Big Data ESSNet web scraping for job vacancies pilot (WP1) during the first nine months of the pilot (from February 2016 to the end October 2016). This is high level summary that describes the main technical challenges and sets out some of the approaches being taken to tackle them. This report discusses some of the work done to date and the issues identified but does not present fully developed methodological solutions. These are still a work in progress.

The report starts with a brief description of job vacancy statistics produced within the European Statistical System (ESS). Details for the participating countries are provided in the Annex. Section 3 describes the high level methodological issues between web scraping from job portals versus enterprise websites. Section 4 then explains the difference between a job advertisement and a job vacancy and some of the issues involved in measuring job vacancies from multiple job portals. Section 5 proposes a conceptual model for evaluating the coverage of on-line job advertisements with respect to the target population. The report then concludes with a summary of work being undertaken within each country participating in the work package and the next steps.

2. Job vacancy statistics within the ESS

Job vacancy statistics within the ESS are current subject to EC regulation No. 453/2008. This defines a job vacancy as:

"… a paid post that is newly created, unoccupied, or about to become vacant:

(a) for which the employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned; and

(b) which the employer intends to fill either immediately or within a specific period of time."[1]

EC regulation 453/2008 has several mandatory elements:

- Quarterly data that has been seasonally adjusted
- Data broken down economic activity (using NACE[2])
- Data is relevant and complete, accurate and comprehensive, timely, coherent, comparable, and readily accessible to users.

[1] Regulation (EC) No 453/2008 of the European Parliament and of the Council http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:145:0234:0237:EN:PDF

[2] http://ec.europa.eu/competition/mergers/cases/index/nace_all.html

There are other elements that are optional, or subject to feasibility, including:

- Job vacancies in the agriculture, forestry and fishing sectors
- Job vacancies in public administration, defence and education
- Data on businesses with less than 10 employees
- Distinguishing between fixed term and permanent jobs.

Member states are granted considerable flexibility regarding the implementation of regulation 453/2008 in the national statistical systems. Some countries use stand alone surveys, while others combine the job vacancy survey with other business surveys. Some collect the minimum information required by the regulation while others collect more. Although the regulation states that the data shall be collected using business surveys, the use of administrative data is equally permitted under the condition that the data are "appropriate in terms of quality" (according to the quality criteria of the European Statistical System).

This has also made it possible to integrate the new European data requirements into pre-existing national job vacancy or business surveys. For example, prior to the regulation, Germany had a comprehensive annual job vacancy survey that ran during the 4th quarter of each year (see Annex, Section 1). Following the regulation, this was supplemented by short follow-up interviews by telephone in the three subsequent quarters in order to meet the requirements of the regulation (Kettner and Vogler-Ludwig, 2010; Moczall et al., 2014).

Further information about the production of job vacancy statistics of each participating country within WP1 is presented in the Annex.

3. Job Portals versus Enterprise websites

As described in the SGA-1 description for WP1, the scope of this pilot is to investigate both job portals and jobs advertised directly on enterprise websites as potential sources of on-line data. The strategy is to focus initially on job portals during SGA-1, and then to switch focus to the approach of collecting data directly from enterprise websites for SGA-2. The general methods for web scraping enterprise websites are being developed by Big Data ESSNet WP2. The intention is to wait until these methods are more developed before applying them to the specific use case of estimating job vacancies.

Although, work on web scraping enterprises has yet to start in earnest, it is worth outlining the main advantages and disadvantages of these two approaches since the methodological challenges are quite different. An understanding these differences can help explain the current methodological focus of this pilot.

For job portals, the key information about a job vacancy (e.g. job title, job location, company name) can be captured relatively easily since this information is usually presented as part of a defined structure[3]. In contrast, extracting any type of structured information from job advertisements on enterprise websites is difficult because the structure of each website is different. However, the approach of using enterprise websites has an important advantage over job portals, namely the existence of an explicit link between the enterprise unit from the survey (or business register) and information about any job vacancies via a website URL. These links are important both for helping to measure the coverage of the on-line data source but may also provide the basis for any estimation approach combining existing survey data with new on-line data sources.

In general, job portals do not use identifiers that would allow them to be linked directly to job vacancy survey data or business registers. Therefore the link between the advertising business and the enterprise unit must be achieved via matching on the business name with other auxiliary information. The experience from this pilot to date is that applying record linkage methods to these data is difficult. One problem is that the legal and trading names of the advertising business are not always the same. In addition complex enterprises there may also be differences the reporting unit. Also, many on-line jobs are advertised through employment agencies and so information about the business with the vacancy is often not contained within the job advertisement. Therefore, the initial methodological challenge with using job portal data is around matching and reconciling these data with existing job vacancy surveys.

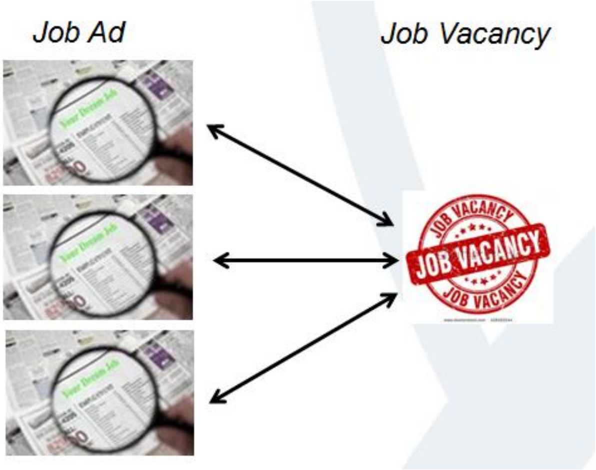
4. Job Vacancies versus Job Advertisements

There is a crucial distinction between a ***job vacancy*** and a ***job advertisement***. The formal definition of a job vacancy has already been defined in Section 2, but in simple terms can be understood as the vacant position within an enterprise. An online job advertisement is defined as the text on a website displaying information about a job vacancy. Thus, the job vacancy is the target concept that we are aiming to measure while an online job advertisement is the target measure that indicates the existence of a job vacancy within an enterprise. However, there is very rarely a one-to-one relationship between the two and this presents some formidable challenges for using on-line information for job vacancy statistics.

Typically, a job vacancy will be advertised on a number of different portals (Figure 1). This may be because an advertisement uploaded on to one portal is republished by others, or because the employer uploads the advertisement multiple times. Thus, deduplication is key challenge for incorporating job portal data into official statistics.

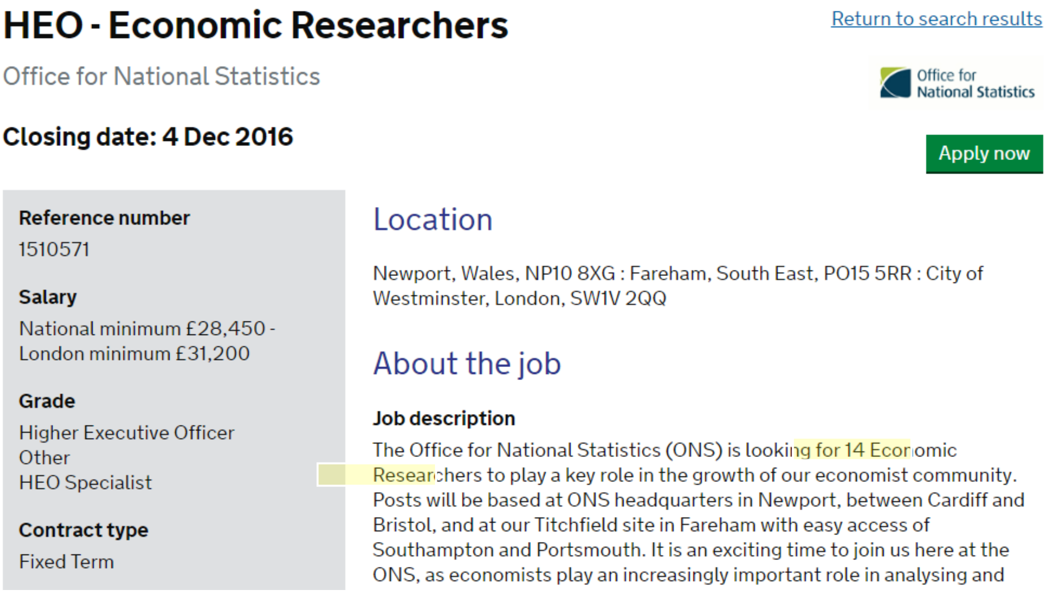
[3] There are considerable challenges in creating structured information from the unstructured elements of a job advertisement, notably the text of a job description which will normally contain information such as skills and education requirements.

Figure 1: Relationship between job advertisements and a job vacancy



It is also possible for a job advertisement to contain more than one vacancy. Figure 2 shows an example of a job advertisement by the UK Office for National Statistic specifying 14 distinct vacancies. In this case it may be possible (albeit very challenging) to develop a methods for parsing job advertisements to extract this information and adjust the count of vacancies. In other cases the actual number of vacancies may not be shown, in which case it might only be possible to flag that an advertisement is advertising more than one position. There may even be instances where an advertisement appears to be advertising a single vacancy, but actually there is more than one. In this case there is no practical solution.

Figure 2: Example of Job Advertisement with Multiple Vacancies



HEO - Economic Researchers

Return to search results

Office for National Statistics

Office for National Statistics

Closing date: 4 Dec 2016

Apply now

Reference number
1510571

Salary
National minimum £28,450 -
London minimum £31,200

Grade
Higher Executive Officer
Other
HEO Specialist

Contract type
Fixed Term

Location

Newport, Wales, NP10 8XG : Fareham, South East, PO15 5RR : City of Westminster, London, SW1V 2QQ

About the job

Job description

The Office for National Statistics (ONS) is looking for 14 Economic Researchers to play a key role in the growth of our economist community. Posts will be based at ONS headquarters in Newport, between Cardiff and Bristol, and at our Titchfield site in Fareham with easy access of Southampton and Portsmouth. It is an exciting time to join us here at the ONS, as economists play an increasingly important role in analysing and

Another issue is that of “ghost vacancies”. These are job advertisements that do not advertise a genuine vacancy. It is known that employment agencies may advertise a vacancy as a means of enticing jobseekers to submit their resumes so that they can increase the number of job seekers they can offer to prospective employers to fill some future vacancy. Although, ghost vacancies are primarily associated with employment agencies, it is believed that employers may also occasionally directly advertise vacancies that do not formally exist. One possible reason is that a company may have purchased a contract for a certain number of postings and so may continue to publish advertisements to use up their allowance - this costs them nothing while they may benefit by having potential future employees on file.

Within this ESSNet pilot, it has been noted that some job advertisements from recruiters follow certain patterns (e.g. “Well known company seeks….”. It is therefore possible that ghost vacancies follow certain semantic patterns that could allow them to be identified and removed.

In conclusion, while there is a clear link between job advertisements and job vacancies, the relation is complicated and producing estimates of job vacancies in not simply a matter of counting job ads.

5. Conceptual Model of Online Job Vacancy Coverage

As well as the challenge of untangling the relationship between online job advertisements and the vacancies they refer to, there is also the important question of the extent to which online job vacancies represent all current vacancies. Although many jobs are advertised online, some may be advertised in other ways, or filled through other means, such as through personal contacts. This leaves a gap between the target population and what can be measured from online advertisements. This gap may vary considerably between countries and by employment sector. For example IT jobs are typically much better covered on-line than jobs in the retail sector.

It is proposed that understanding of these gaps is key to understanding how online job advertisements might be incorporated into official statistics. Further, it is assumed that for most (and maybe all) countries the gap between online job coverage and the target population is such that it would not feasible to produce estimates of the target population based purely on on-line job advertisements. Therefore, it is envisaged that job vacancy surveys will need to continue in some form to provide a benchmark to which data from other sources would be calibrated. Other sources would be incorporated to provide additional variables that are not captured from surveys and possibly more frequent and timely estimates. Therefore, the main methodological challenge for this pilot is around combining and integrating data from different sources.

Figure 3 proposes a conceptual model of how online job advertisements correspond to the target population. In practical terms this may be defined as all vacancies that are available to be measured by existing job vacancy surveys. This model assumes a “clean” data set with all duplicate advertisements removed and adjustments for advertisements offering multiple vacancies. All relevant information would be reduced to a single set of variables for each on-line job vacancy. Other features of this model include:

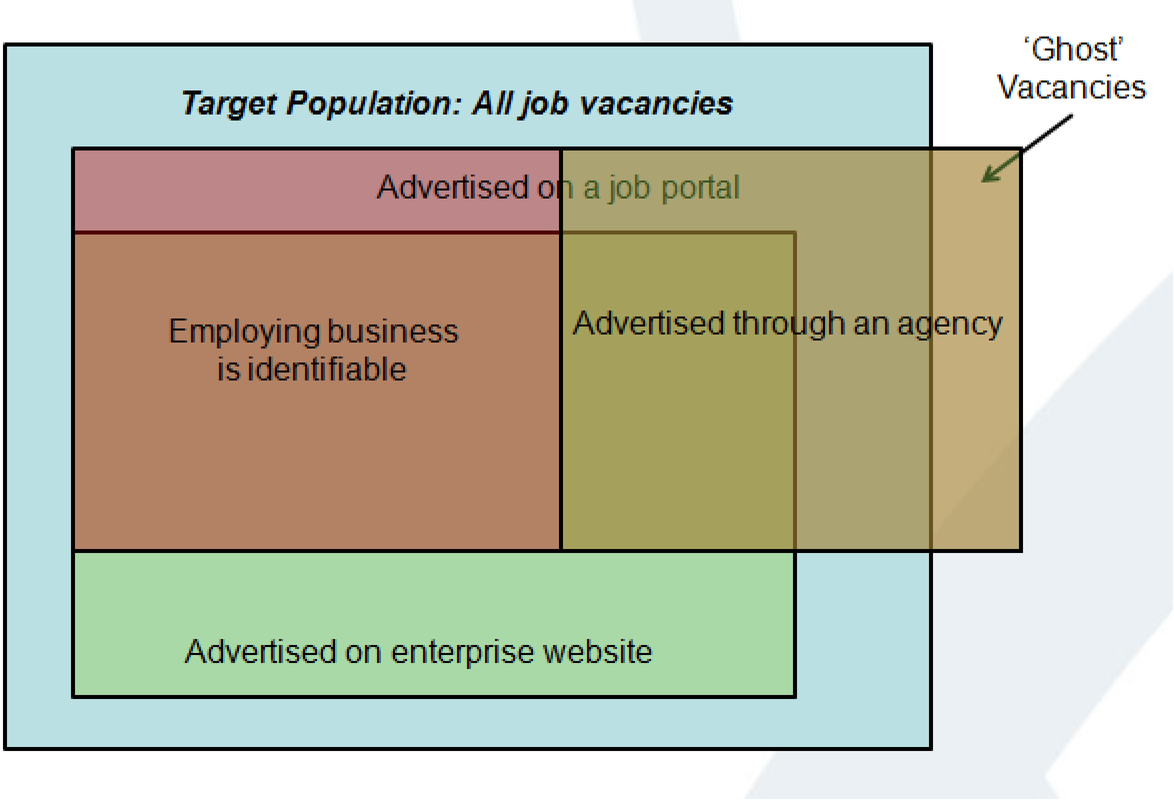
i. The target population corresponds to what is estimated by job vacancy surveys.

ii. Jobs advertised through online portals include both jobs where the employer is identifiable and jobs that are advertised through an employment agency where the employer can not be identified. There are specific challenges with the latter in matching these to enterprise units from the vacancy survey or business register.

iii. Jobs advertised through online job portals may also include some ‘ghost’ vacancies that are not within the target population. The model shows this as a subset of jobs advertised by employment agencies although a small number may also exist where the advertising business is identifiable.

iv. Jobs advertised directly on enterprise websites are a subset of the target population and will largely (although not completely) overlap with job vacancies advertised on job portals.

Figure 3: Conceptual model for measuring job vacancies from on-line sources.

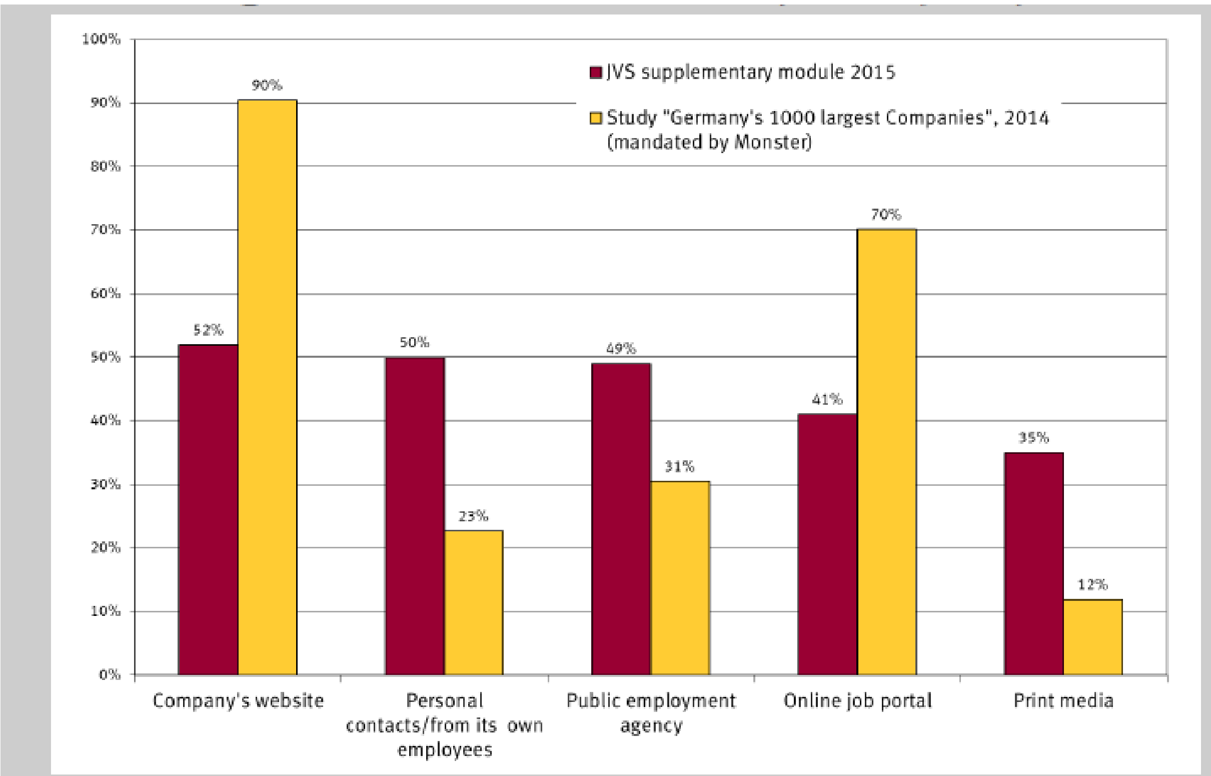

Target Population: All job vacancies
'Ghost' Vacancies
Advertised on a job portal
Employing business is identifiable
Advertised through an agency
Advertised on enterprise website

Current job vacancy statistics are subject to various sources of error such as sampling error, respondent error, and non-response bias. There may also be issues for newly created businesses who are advertising vacancy since such businesses may be underrepresented on the register. However, the working assumption is that the job vacancy estimate based on the survey is closer to the "true value" of the number of job vacancies than the unique number of jobs advertisements.

As well as providing a conceptual framework for understanding the coverage of job vacancies from online sources and how these relate to the measurement of all job vacancies, this approach may also provide the conceptual basis for an estimation framework. The details of this are sketchy at present, but might involve an initial matching or reconciliation of enterprise level data between on-line sources and the job vacancy survey followed by modelling of residual data that cannot be directly matched, including jobs advertised through employment agencies.

Other information that would be useful within this kind of estimation framework is data on the recruiting channels used by enterprises. Germany already has some of this information collected through their recent annual JV survey along with equivalent information for large enterprises from a separate study (Figure 4). This shows that larger enterprises are more likely to use on-line channels whereas small businesses are more likely to use traditional channels, such as print media. Slovenia are also actively considering incorporating new questions about advertising channels on their job vacancy survey, to provide additional information on recruitment channels. France (who will be joining for SGA-2) is considering something similar. This kind of data would help with understanding the differences between those enterprises that advertise on-line and those that don't, or primarily use other channels. In turn, this information could be used for producing hybrid estimates that were partially based on directly on-line job advertisements and partially on model-based estimates.

Figure 4: Recruiting Channels used by German Enterprises



© Thomas Körner and Dr. Martina Rengers, Federal Statistical Office, labour market

6. WP1 Progress Reports by Country (February 2016 to October 2016)

Each participating country in this pilot is faced with a unique set of circumstances. There are major differences in terms of population size/number of job vacancies, the number of job portals, the relative importance of on-line channels for advertising job vacancies, the availability of other sources (including access to job vacancy survey data) and each NSI's legal position on web scraping. There are also differences in terms of domain expertise, access to technology, and previous skills in working with big data. These factors mean that the approaches being taken and the rate of progress is different for different countries. For example, Slovenia have already made significant progress with the reconciliation work outlined in Section 5 including even data collected from enterprise websites. However, Slovenia has the advantage of being a small country with only two main job portals and relatively small volumes of data.

The approach being taken by WP1 is to develop approaches that are relevant to the specific context of each country in terms of accessing, processing and analysing data but to share knowledge and experiences and identify common approaches where possible. The need to follow different paths is beneficial as this increases the chances of identifying new ideas that we have might not have otherwise considered.

Despite these different approaches, there are a number of common methodological challenges. In an effort to develop common approaches, the pilot held two 'virtual sprints'. This involves WP1 participants from different countries working on a common problem at the same time and sharing experiences and results. The first virtual sprint was held of 28-29 July 2016 and focused on the subject of deduplication. Although this produced some promising approaches[4], we concluded that achieving high quality results would require a lot of investment in time and that we would be better focus on the problem of matching job portal and survey data as this is more central to answering the central question of the coverage of job portal data against what is measured by the job vacancy survey. A second virtual sprint was held on 29-30 September and focused on this problem[5]. This sprint also made some progress although it is clear that matching the advertising business on a job advertisement with an enterprise unit is a difficult problem.

The remainder of this section provides further details of progress being made within each country.

6.1 Germany

6.1.1 Inventory of job portals

The project started with an inventory of job portals in Germany that was to serve as a basis for the selection of job portals for web scraping (Koerner, Rengers et al. 2016). The study revealed that in Germany, the number of job portals is particularly high, varied and changing dynamically. There are more than 1000 job portals that exhibit important differences concerning the business models used by the job portal owners. Three different types of job portals must be distinguished: job boards, job search engines and hybrid portals. While job boards are directly commissioned by employers to host their job advertisements, job search engines reproduce job advertisements that were originally found at job boards. Hybrid job portals combine both approaches. All this leads to the fact that duplicate job advertisements can be found within and between the job portals. Web scraping job portals can therefore make de-duplication procedures vital.

A qualitative assessment of the information available on job portals shows that the richness of the information available in structured form tends to be rather limited on many job portals. Most job portals return a summary list of search results, which only shows a limited range of variables available in structured format (usually included the job title, the employer, the location and the date of the advertisement).

A further aspect concerns the dynamics of the job portal environment. It is difficult to provide a detailed account of these changes. In large countries, such as the UK and Germany, the number of job portals is too vast and dynamically changing to undertake a comprehensive overview. However it is important to have an understanding about the speed of changes as such changes may require changes in the selection of the job portals.

[4] Notes from the 1st virtual sprint are documented on the ESSNet Wiki: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP1_2016_07_28-29_Virtual_Sprint_Notes

[5] Notes from the 2nd virtual sprint are documented on the ESSNet Wiki: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP1_2016_09_29-30_Virtual_Sprint

For the further development work in the project, based on an assessment of the job portals included in the inventory, it was decided to focus on four major portals in Germany that would have to be included (job portal of the Federal Employment Agency, Gigajob, StepStone, Monster).

6.1.2 Data acquisition

The data acquisition was preceded by a study of the legal framework for web scraping. The two most relevant legal areas are the terms of use of the job portals (often more or less explicitly forbidding the use of web scraping technology) and the copyright laws. In both cases the situation was found to be ambivalent: It is a complex legal question whether the terms of use of a specific job portal would apply to a statistical office or could rather be considered to be invalid in this situation. Similarly, the copyright protection of data bases is a relatively new and difficult topic to which a clear cut answer is not easy to find. In short, one may state that the lesser the share of information obtained from a database, the smaller the likelihood of legal concerns to web scrape this information. As a result, it was concluded the legal restrictions were no major issue for the pilot study, but that for any larger scale production of statistics it would be highly recommended to obtain the consent of the portal owner, not least to avoid breaks in time series due to lacks in data availability.

As part of the project, we plan to study whether the data owners of important portals are ready to cooperate with official statistics. A first contact has been established with the Federal Employment Agency (FEA) that hosts the largest and most important job portal in Germany. In a one-day workshop on 17 October 2016, the data used by the FEA, in statistical reporting as for their placement services, were discussed in detail with 10 experts from different sections of the FEA in order to identify possible areas of cooperation.

For the analytical purposes of the project, web-scraping technology was applied nevertheless, in order to obtain smaller scale data for the analyses carried out in the project, but also to test the feasibility of web-scraping in the situation in Germany (characterised by large volumes of data).

6.1.3 Deduplication

An exploratory study on deduplication was carried in this context of the first virtual sprint in late July 2016. The main objective of the study was to explore the prevalence and detection possibilities of duplicates on and in between job portals. The approach was an exploratory one: For a subset of data obtained by web-scraping, the prevalence of (possible) duplicates was assessed manually. In suspicious cases, the relevant full job advertisements were checked to investigate whether there actually was a duplicate. The results of the exercise were subsequently analysed regarding the possibilities to define rules for web-scraping based on the structured information.

The results of the sprint showed that, while practically no duplicates were discovered at the job board selected for the test (Stepstone), there was evidence that many duplicates can be found at the hybrid job portal selected (Gigajob), which apparently runs only limited deduplication procedures. Furthermore, it was concluded that deduplication is hardly feasible on the basis of the structural information given in the hit lists of the job portals studied in the sprint (neither inside Stepstone nor between Stepstone and Gigajob). A reliable deduplication necessarily involves considering the plain text of the job advertisements in addition. When including hybrid portals, it should be considered to make only use of “own” job advertisements to reduce deduplication issues.

6.1.4 Comparison of distributions in online job portals and the job vacancy survey

The comparison of the structure of job advertisements found in a major job board with aggregated results of the German Job Vacancy Survey was the objective of the second virtual sprint in late September. Again web-scraped data from Stepstone, the biggest job board in Germany, were used for the test. The scraping was done twice, first 7 September 2016 and second time on 4 October 2016. These results were compared with German Job Vacancy Survey results published for the first quarter of 2016. Although the reference period is not the same, it was found that the structure of the job vacancies with regard to the economic activity of the enterprise is quite stable over time – at least during the year.

Stepstone has the specificity that it includes structured information on the economic activity of the enterprise, which, at least at first sight, seemed similar to the standard classification used in official statistics, the Statistical Classification of Economic Activities in the European Community (NACE). For this reason, compared with other portals that do not provide any structured information on the economic activity of the enterprise, it was the most suitable basis for a first analysis of the structural differences between job advertisements on a major job board with the JVS results.

The results of the sprint showed a number of problems that led to the conclusion that the structured information on the economic activity provided by Stepstone (and possibly also the other portals) is not suitable for statistical purposes, since the sectors used by Stepstone cannot be allocated one-to-one to the NACE sectors. Furthermore, no information is available how the coding is done at Stepstone. Another issue was that the web scraping of the structured information on the economic activity can only be implemented via the use of filters in step stone's advanced search. For unknown reasons, the sum of the job advertisements in all sectors is smaller than the number of job advertisements found without the application of a filter. At the same time, some of the job advertisements were allocated to more than one sector.

Despite the fact that these issues make it almost impossible to draw meaningful comparisons, it is nevertheless quite obvious that the jobs advertised at Stepstone have a different distribution than the German job vacancy survey: The share of jobs in the sector "information and communication" and "financial and insurance activities" (as well as possibly also "manufacturing") is higher at Stepstone, while the share of advertisements in construction sector (and possibly also in the service sectors) is lower.

As the classification used by Stepstone is not an option for statistical purposes, it remains to be investigated whether information on the economic activity can be obtained via matching the business register with the job portal data or by means of a textual analysis of the full job descriptions found at the job portals.

6.1.5 Next Steps

The main focus of the next months will be to further investigate the possibilities to get access to the data from the owners of the most important job portals identified in the inventory. Apart from investigating the readiness of the portal owners to provide their data for statistical purposes, it also needs to be analysed whether further structured information (not available to the general public via the websites) could be made available to enrich the analytical potential of the data. As the information on the economic activity sector, a quite crucial variable for job vacancy statistics, cannot be obtained directly from the portal, another priority is to investigate the legal, technological and methodological prerequisites for the matching online job portal data with business register data.

6.2 Greece:

6.2.1 Data acquisition:

As a first step of the current research, an internet investigation of job portals, job search engines and specialist job sites, was carried out across Greece. According to http://www.greek-sites.gr, which is a site that ranks the Greek sites popularity, 28 job portals with domains ".gr" were found. This list is presented in the inventory Annex (Körner, T., M. Rengers et al., 2016).

To determine which web sites should be focused on within the pilot study, the major job portals are sorted based on the following criteria: a) the number of advertisements (size); b) monthly visitors (June 2016) and c) the Alexa[6] popularity ranking. Taking under consideration the results included in the inventory (Körner, T., M. Rengers et al., 2016), ELSTAT mainly focused on collected data from Kariera.gr and Skywalker.gr.

The ads were scraped directly from job portals using a web scraping tool. An automated "point and click" tool for general scraping purposes (import.io) was used, at the beginning. However, as experience was gained, another web scraping tool (Content Grabber) was chosen, which could wait for selectors, navigate through multiple pages, develop constrains and logic, handle error cases and create "scraping agents".

"Scraping agents" are tailor made programs (scripts) which can be built either by training the tool to collect the data in a specific way and/or by writing scripts using a provided build-in script editor. For each job portal, some key variables such as the job title, location, company name, posted date, salary and job type (full time/temporary) and a "snippet" of the job description between 40-60 words were collected.

6.2.2 Deduplication:

Job portal web sites provide a list of all advertisements which only shows limited range of variables, usually job title, employer's name or logo, location and date of advertisement. Further information could be obtained by using search filters offered by the job portal, that provide the opportunity to select for example job posts for each job category, full-time jobs etc. So, one could start web scraping a portal's web site (startup url), after using a search function.

[6] The Alexa ranking is a well-known metric based on the web traffic data collected by the California-based company Alexa Internet, Inc., a subsidiary wholly owned by Amazon.com.

Although, the information on the startup url looks good, the algorithms used by the portals, very often show the same job advertisement in more than one different categories. For that reason filters used by job portals produce duplicate ads.

In the context of the first virtual sprint the aim was to identify duplicates within each portal. The key to avoid including duplications in the dataset is to use the appropriate URL address of the job ad pages.

The data were scrapped and the dataset for each one of the two job portals was produced. This approach worked well for each portal but more work is needed to examine the removal of duplicates in the common dataset of scraped data from the two portals, at the same period of time.

6.2.3 Matching:

In the context of the second virtual sprint, the experiment undertaken by ELSTAT aimed at exploring to what extent the job portal data covers what is measured by the job vacancy survey. A “clean” sample dataset of 3060 single advertisements was created from the scraped ads collected between 15.6.2016 and 15.8.2016.

The first attempt was to match the employing company names from the description of the job advertisment to the companies from the sample of the job vacancy survey. This approach revealed many challenges. In many cases the employing companies use a trading name different from the “official” name recorded in the survey data. In other cases, ads didn’t include any company name. It was observed that these particular ads usually started by stating “Leading Company…”; or “Well Known Firm…” etc.

In the current experiment, the identification of the employing company names from the description of the job advertisements was possible for only 55% of the ads. For the rest 45% of ads, it was not possible to identify the company names. In detail, no identifiable company names corresponded to the 77% of these ads. Further work is needed to understand in what extent these ads concern “ghost vacancies”.

Next step was to match the 256 identified companies with the ones from the sample of Job Vacancy Survey. Only 9% of these companies were matched. To further explore the coverage issues, the Statistical Business Register was used instead of the sample of J-V survey to match the companies. The results were better in this second attempt. The matching percent increased to 30% of companies.

The matching results of ads and companies, as well as their percentages at the level of the alphabetic code of NACE Rev.2 are presented in the following table.

Table 1: Ads and Companies classification by Economic Activities (NACE rev.2)

Economic Activities Rev.2 Description		Ads	% of Ads	Companies	% of Companies
C	MANUFACTURING	54	11.0%	15	19.5%
G	WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES	74	15.1%	20	26.0%
I	ACCOMMODATION AND FOOD SERVICE ACTIVITIES	23	4.7%	5	6.5%
J	INFORMATION AND COMMUNICATIONA	20	4.1%	4	5.2%

K	FINANCIAL AND INSURANCE ACTIVITIES	21	4.3%	5	6.5%
M	PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES	153	31.3%	9	11.7%
N	ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES	18	3.7%	6	7.8%
P	EDUCATION	74	15.1%	6	7.8%
Q	HUMAN HEALTH AND SOCIAL WORK ACTIVITIES	20	4.1%	2	2.6%
S	OTHER SERVICE ACTIVITIES	24	4.9%	1	1.3%
D,F,H,R	OTHERS	8	1.6%	4	5.2%
		489	100.0%	77	100.0%

These preliminary results show that almost one out of three job advertisements (31.3%) are classified in the branch of economic activity M (Professional, Scientific and Technical Activities). Moreover, almost 46% of the companies, which advertise online job vacancies, are classified in the branches of economic activity G (Wholesale and retail trade) and C (Manufacturing). Further work is needed to study the coverage issues.

6.2.4 Next Steps

As mentioned above, ELSTAT will focus on continuing the work of matching the job portals data to Job Vacancies survey data and to Statistical Business Register, in order to better understand the issues of coverage.

6.3 Slovenia

6.3.1 Assessment of job portals:

In Slovenia 9 job portals and 109 job agencies have been identified. There are two main job portals (Moje Delo, Moja Zaposlitev), which together publish 95% of all job vacancies advertised on portals.

Import.IO is the main tool used for scraping. APIs have been run manually every Monday morning from mid-May onwards. When survey on job vacancies collects data (on reference day), we scrape data on the reference day and also the following day. Job vacancies that were valid on reference day are those scraped on the first day and were published during that day. Additionally, if a job vacancy is published for more than a month, this is assumed to be an invalid job vacancy.

During the collection period, each job portal scraped two ways: First a list of all job vacancies on the domain, and second we scrape a content of specific job vacancies. The following variables are collected:

- Job vacancy title
- URL address of job ad page
- Company
- Place of work
- Date of published of the job ad

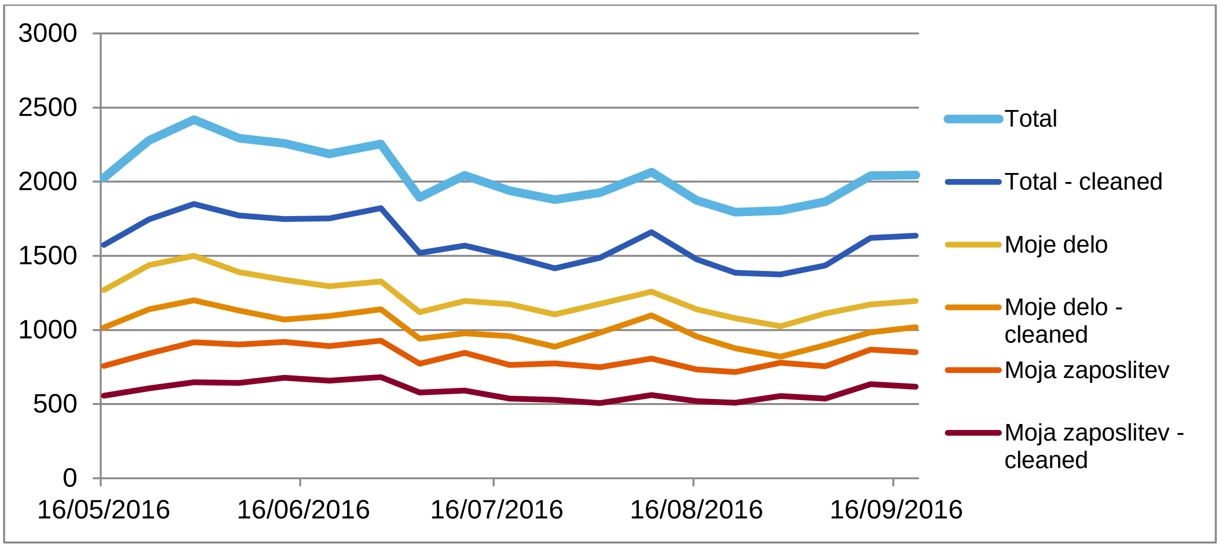
- Description of job vacancy
- Description on how and until when to apply for the job (only on Moja Zaposlitev)
- Date of scraping (not scraped, but created)

The weekly volume is between about 2,000 to 2,400 queries depending on how many job vacancies are published.

6.3.2 De-duplication

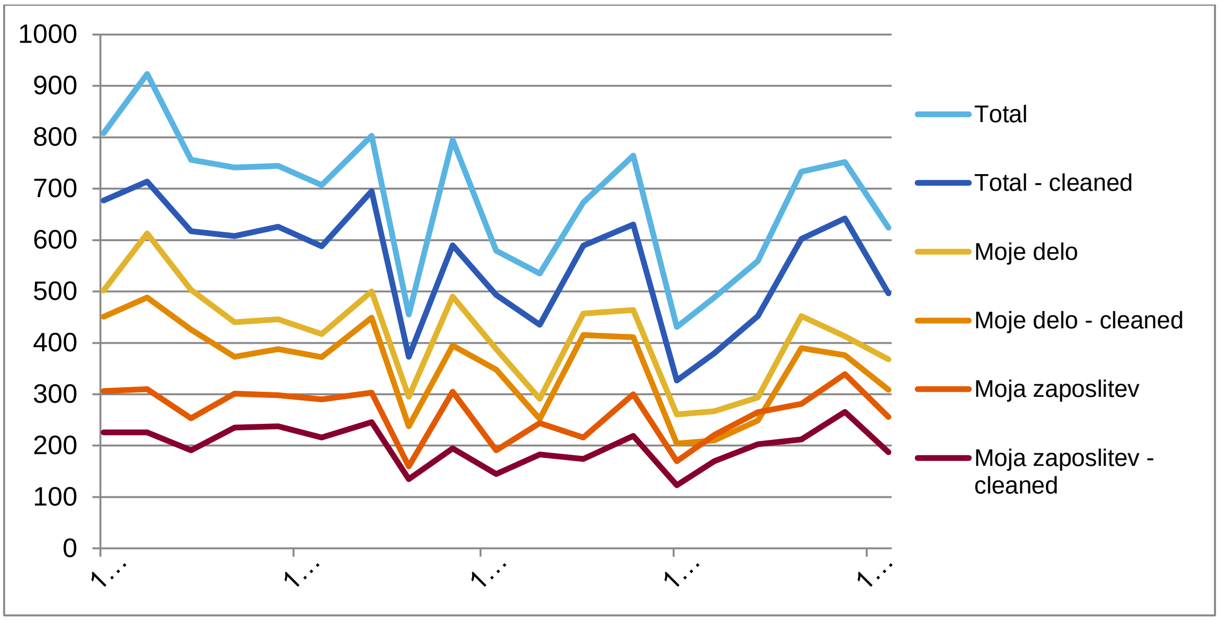
Job ads are scraped every week and so some vacancies are scraped more than once. When creating monthly data sets, we eliminate observations that are detected multiple times. The key for removing duplicates is URL address of the job ad page. Between 19% and 24% of job ads are for working abroad or for unpaid student positions. These are removed as these are not included in the target population. The level of all job vacancy ads scraped every week from mid-May to mid-September is shown in Figure 8. The "cleaned" line represents data without ads for student work and work outside of the country. The trend is the broadly the same whether we look at the raw or cleaned data set and is also similar for each portal.

Figure 5: Weekly counts of job vacancies by portal (All)



The second graph shows only "newly" published job vacancy ads – ads that were published from one week to the other. Again, the pattern is similar no matter if data set is cleaned or not and regardless the job portal. However the trend is not as smooth as if we look all the scraped ads.

Figure 9: Weekly counts of job vacancies by portal (New jobs only)



6.3.3 Record linkage with The Slovenian Business Register

After removal of redundant ads we combine scraped data with The Slovenian Business Register using record linkage techniques. The main variable used for linking is company name. The aim of this task is to assign a registration number of a company to the job vacation ad. We are able to do that for roughly 99 % of the job vacancy ads.

First step is to clean the data. All words are converted to upper case, and special characters are removed (e.g. .,&+-*/). Some companies do not have unique names and sp we create two data sets: one with unique company names and one with duplicate company names.

The company name from job ad with short name/full name/abbreviated name of a parent company from The Slovenian Business Register. This is done in 5 different ways in 13 steps.

We are able to link about 90% of job vacancy ads in first 3 steps; that is with direct comparison of unique company name from job vacancy add to short name/full name/abbreviated name from the Register (steps 1.1, 1.2, 1.3). Some company names in the register include the headquarters location. If so, these are deleted and the direct matching step is repeated. In this steps (2.1, 2.2, 2.3) we link about 2% of job vacancy ads.

If a company does not have a unique name, a reliable match cannot be made on the name alone. Therefore, when directly linking data with a data set containing duplicate company names, we choose that observation where place of headquarters is the same as the location of the advertisement. This tactic is only efficient when comparing the short name from the Register (step 4).

An additional 2.5 to 3% of job vacancy ads are linked by calculating the distance between the strings. Function, that determines the likelihood of two words matching, expressed as the asymmetric spelling distance between the two observations, is being used. If the distance is 5 or less, records are written in a new file. The right matches are then determined manually. This is done in step 7 (short name) and step 8 (full name). The same method is used in step 11, but the distance here is 20 and we are using only full name.

Some companies publish a lot of job vacancies, but it was not possible to link them using the above steps. Therefore we manually searched for the proper registration number. A list of those results was made and was updated after every linking new scraped data. This way we are able to link about 3 % with "old" list and additionally we manually search and find another 1.5 % of job vacancy ads. This is done within step 10.

The structure of record linkage by steps is the similar no matter the period of data or the amount of scraped ads. The full results are shown in Table 2.

Table 2: Matching Results from Slovenia JV survey to Business Register

Step	**31.5.**		**31.8.**		**May**		**June**		**July**		**August**	
	N	%	N	%	N	%	N	%	N	%	N	%
1.1	1401	72.03	1271	75.61	2109	73.28	2662	72.59	2192	73.95	2450	75.18
1.2	365	18.77	250	14.87	488	16.96	645	17.59	506	17.07	510	15.65
1.3	3	0.15	2	0.12	3	0.1	4	0.11	1	0.03	7	0.21
2.1	22	1.13	24	1.43	46	1.6	54	1.47	27	0.91	29	0.89
2.2	9	0.46	13	0.77	16	0.56	21	0.57	16	0.54	24	0.74
4	11	0.57	15	0.89	20	0.69	33	0.9	30	1.01	27	0.83
7	16	0.82	27	1.61	28	0.97	31	0.85	30	1.01	26	0.8
8	11	0.57	9	0.54	16	0.56	20	0.55	8	0.27	25	0.77
10	82	4.22	49	2.91	100	3.47	109	2.97	83	2.8	98	3.01
11	17	0.87	2	0.12	21	0.73	29	0.79	19	0.64	15	0.46
0	8	0.41	19	1.13	31	1.08	59	1.61	52	1.75	48	1.47
TOTAL	**1945**	**100**	**1681**	**100**	**2878**	**100**	**3667**	**100**	**2964**	**100**	**3259**	**100**

6.3.4 Removal of duplicates

Some companies publish an advertisement for the same job vacancy on multiple job portals. We consider a job vacancy ad as a duplicate if it is published on a different portal and if it matches with some other ad on the register number of company, on job vacancy title and on place of work. Ads are considered 100 % duplicates if the frequency of ads by register number of company, job vacancy title and place of work is the same on all portals. If it's not the same, we consider these additional advertisements as duplicates so we do not count them.

Regarding duplication, we analyzed job vacancies ads data that were current on the reference day. The results are produce in Tables 3, 4, and 5.

Table 3: Number of companies by job portal

Code	31.5.		31.8.	
	N	%	N	%
0	817		818	
0.1	518	63.40	538	65.77
0.2	241	29.50	221	27.02
0.3	63	7.71	59	7.21

Code 0 in Table 3 represents the number of all companies that published a job advertisement on one of the two largest job portals. The table shows that about 65% of all companies that had job vacancies on the reference day had a job vacancy ad published on our largest job portal (code 0.1) and about 30% had published on the second largest portal. There were about 7% of companies that had job vacancies ad published on both portals.

Table 4: Number of companies by uniqueness of published ads

Code	31.5.		31.8.	
	N	%	N	%
1	563	68.91	536	65.53
2	285	34.88	254	31.05
3	42	5.14	47	5.75
4	5	0.61	2	0.24

Most of the companies, 65 per cent, published unique ads on our largest job portal (code 1), while a third of them published unique ads on our second largest job portal (code 2). A fraction of them, that is 5 %, publish duplicate job vacancy ads on both of our largest job portals (code 3), while less than 1 % of companies published job vacancy ads for which we cannot be completely certain whether they are duplicates (code 4).

Table 5: Number of job vacancy ads by job portals

Code	31.5.		31.8.	
	N	%	N	%
1	825	46.56	632	41.63
2	410	23.14	398	26.22
3	57	3.22	66	4.35
4	3	0.17	6	0.40
5	447	25.23	416	27.40
Total	1772		1518	

About 40% of job vacancy ads were published only on our largest job portal (code 1), while about quarter of them were published only on our second largest job portal (code 2). There were about 4% of duplicate job vacancy ads – ads for the same job vacancy that were detected on both job portals (code 3) – while less than 1% of ads are assumed to be duplicates (code 4).

Code 5 represents job vacancy ads published by job agencies representing about a quarter of all job vacancy ads.

6.5 Sweden

The legal basis for Statistics Sweden to do web scraping is not clear and no data has yet been captured using web scraping. Statistics Sweden thus decided to concentrate on job portals, and in particular data from the portal of the Swedish Employment Agency (SEA), "Platsbanken" (PB). SEA is a main player on the labour market in Sweden and provided us with data directly from their database. The approach taken is to explore the PB data as a starting point to understand job portal data, and to evaluate the quality of the PB data.. When we know that the data quality is good enough, linking to the job vacancy (JV) survey and the Business Register (BR) is relevant.

The next step will be to make contact with other (private) job portal owners on the inventory list generated in Deliverable 1.1, and acquire data from them. It is anticipated that these data could be used as a complement to the PB data. The methods tested and evaluated using the PB data will help the analysis of other portal data.

A deduplication study will be carried out during the next step, partly because the findings in the first virtual sprint indicate that sophisticated methods need to be developed, and partly because such a study will be more relevant when multiple data sources are involved.

6.5.1 Data description

The PB data contains all advertisements between January 1st 2012 and June 30th 2016, in total 2 206 072 job advertisements. The data is structured and contains 20 variables, e.g. date posted, date for final application, company identification number and name, municipality, job category and occupation, and the titles and the texts of the advertisements. The company identification number, the occupation identification number, and the municipality code use standard code systems, which makes it easier to link to other data sets.

6.5.2 Data exploration

The data were processed with Python, and stored in a Microsoft database. The characters, coded or not coded, are defined in varchar[7] of different lengths. Data types Date and Integer are unchanged in our database. MS SQL server and R were used for processing the data.

The data were evaluated by a few simple generic quality indicators. This is important in order to get an understanding of the basic data quality when matching to the JV survey or in further analysis. For example, if organization identification number is used for searching, grouping, and linking to other

[7] A data type for handling text data.

datasets, the value should be available for most of the advertisements, and the value should at least be numbers of the correct length.

Table 6 presents the valid value percentages of some important variables. If the value is null, 0, empty space, non-digital, or non-characters, it is an invalid value. The date the advertisement was posted and the final date for application should fall between January 1st 2012 and August 31th 2016. The high percentages of valid values of these variables indicate that they can be useful for further analysis. Variables with large percentages of missing values and error values need to be improved, if possible, before they can be used or they should not be used for further analysis.

Table 6: Percentage of valid values of the main variables

Variable	*Percentage of valid values*
Organization id	99.2%
Company name	~100%
Municipality code	97.96%
Occupation id	99.54%
Employment category	~100%
Title of advertisement	~100%
Description of advertisement	99.91%
Number of job openings	84.13%
Date of posting	100%
Final date for application	98%

Most advertisements are posted by recruiting and outsourcing companies, as showed by the organization id and the name of the company. The six companies with the largest number of advertisements are such companies; their share in the PB data is about 4,2% (0.91%, 0.86%, 0.80%, 0.59%, 0.53%, and 0.48% respectively) during the entire period of four and a half years.

The analysis of the municipality shows that advertisements are mostly for jobs in the large cities. Stockholm, Gothenburg, Malmö, and Uppsala are the most featured municipalities.

The high skilled jobs are well covered in the PB data base, as showed by the occupation id and when comparing with main occupation categories. Two categories of jobs i.e. *plant and machine operators* and *assemblers and elementary occupations* are completely missing from the advertisements. In general, low skill jobs seem to be less advertised on the portal, compared to high skill jobs.

It is also interesting to see for how long a job advertisement is posted on the portal. This information is important when matching to the JV survey (see below). The period of posting must cover the reference day of the survey for an exact match. For the duration of days, we use the final date for application minus the date the advertisement was posted. We cannot be sure that the post is removed immediately when a job is no longer vacant, so this is an assumption made. The mean number of days is 26, and the median is 21 days. 50 per cent of the advertisements stay on the portal between 14 to 30 days. Advertisements with a difference larger than 360 days were not used in the calculation (561 advertisements).

6.5.3 Data matching

The PB data were matched to the BR by matching on the organization identification number. Using the matched data, the company names in the PB data and the BR were compared. Municipality and the occupation identification were matched to the standard Swedish municipality code and the SSYK2012 (compliant with ISCO-08). The results are shown in Table 7.

Table 7: Variable matching results

Variable	*Percentage of matched values*
Organization id	99.02%
Company name	0.28% (0.01%)
Municipality	97.96%
Occupation	6.07%

99% of the advertisements find matches in the BR. Of the less than 1% that do not find a match, this is either because the organization ids are lacking in the PB data base, or there are no matches in the BR. However, we suspect that many companies that have more than one work location post advertisements with the same organization id. For those matched companies, the company names used in the PB were compared with the registered names of the enterprise legal units. The string match algorithms jaro-winkler (jw) and longest common subsequence (lcs) were applied.

The overall matching rates are very low. With the threshold jw=0.7 and lcs=70, the matching rates are only 0.28% and 0.01%. In future work, we need to compare further with other names in the BR, e.g. names of the enterprise work units.

For example, “DISTRIKT UPPLAND, MAX I UPPSALA 2 – BOLÄNDERN”(company name in the PB) and “MAX HAMBURGERRESTAURANGER AKTIEBOLAG” (company name in the BR) do not match, because the business register saves the officially registered name, but it is a work place in Uppsala that posted the advertisement. There are also other examples, such as “SVENSK MILJÖISOLERING AB” and “DEX-TERA H-T DYNAMIC CONSTRUCTION SOLUTION SKANDINAVIA AB”. These names are totally different but refer to units with the same business identification.

There could be several reasons for the mismatches:

1. The organization id alone as the linking key is not accurate enough, other variable such as municipality and address with good quality need to be added for linking;
2. The company names used in the PB database are not controlled when an advertisement is posted, and it is possible to add text such as name of a city;
3. There are other types of company names in the BR that may be more reasonable to use for matching.

Matching the municipality code to the regional code gives 98% correct matches, and these will be further used to match businesses and work places. The occupation is matched on a detailed level of the SSYK 2012. 6.07% match is very low considering that Table 1 shows that only 0.46% do not have an occupation id, which means that most of the advertisements have an occupation id that should match to the standard. However, the reason could be that an older standard is still in use, and results might improve using the older standard for matching.

With the preliminary assumption that the organization id provides reasonable matches with the BR, the coverage of the enterprise sector is examined in order to get a first indication of how the PB advertisements are distributed compared to estimates from the JV survey and the distribution of businesses in BR (Table 8).

Table 8: Distribution of businesses by sector and data source

Sector	*PB*	*JV survey*	*BR*
Private	69.65%	91.78%	89.70%
Non-profit organization	1.34%	1.06%	10.24%
Public	16.87%	7.00%	0.05%
Other	11.16%	-	-

The percentages for PB in Table 3 are calculated based on the total number of advertisements. If few companies post many advertisements during the period, their frequent appearances contribute to the percentage of the group. Advertisements in PB cover both large and small companies, although the big companies have the highest percentage. This is reasonable considering that bigger companies need a larger work force. Table 3 shows that all sectors are found in the PB data. Null means that some enterprises lack sector code. The coverage of the economic activities NACE code is also very high, about 94%. The differences between the PB and the JV survey need to be further explored. We know that the PB does not cover all types of occupations (e g less low skill occupations). The difference in the public sector is interesting. Comments from the management of the JV survey give at hand that there are suspicions of under coverage of the public sector, and if this is the case, the PB might be used to support this part of the JV survey. However, before any conclusions can be drawn about the possibility to use PB for official statistics, much work remains to be done.

The PB data were matched to businesses participating in the JV survey, covering the same period January 1st 2012 and June 30th 2016, 410 393 responding units in total. Legal units are sampled for the public sector while work place is the sample unit used for the private sector (see the Annex for further details). When PB data are linked to survey data from the public sector, we use organization id, year and month of advertising; the combination of variables of organization id, municipality, and year and month of advertising are used to link to the private sector. 71 % of the sampled records in the public sector find a match in the PB. Only 14% of the records in the private sector could be matched to the PB data. The low percentage of the match can depend on the keys, i.e. it is not enough to identify the enterprise work unit with the organization id and the municipality. This will be further investigated.

When matching to the JV survey, our main interest in the first step was to what extent the same businesses appear in both the survey and the PB database. We have not yet studied to what extent advertisements might be duplicated by the businesses appearing in both JV and PB, this will be taken care of in future work.

6.5.4 Future work

We have gained a first overview of the job portal data and its quality. The job portal data provide rich information and also challenges pertaining to quality and linking. For the remaining part of SGA 1, we will continue to explore the quality of the PB data. Our matching strategy will be improved and methods for deduplication will be investigated. We will further investigate how the PB data can be used to inform the current job vacancy survey. Questions like how to handle the large number of advertisements posted by recruiting companies will need to be considered. We will also investigate text mining methods on the full text of the advertisements to evaluate and improve the PB data quality. Contacts will be made with other job portals and we will hopefully get additional data from these sources.

6.6 United Kingdom

6.6.1 Data acquisition

The UK has focused mainly on collecting data directly from several large UK job portals (including Indeed, Adzuna, Totaljobs, and Careerjet). The primary method is through the use of APIs provided by the portals, which has enabled access to quite large amounts of data fairly easily.

6.6.2 Deduplication

An experimental approach for identifying duplicate job advertisements has been developed based on Dedupe[8], a Python library specifically designed to support the identification of duplicate records within a list. This uses a supervised learning approach to train a logistic regression algorithm to

[8] https://github.com/datamade/dedupe

recognise duplicates based on clerical decisions about whether two job advertisements relate to the same vacancy.

The experiment involved collecting job advertisements in the IT and construction sectors from three large portals over a week. The first main processing task was to produce a standardised set of variable names to be used for matching (i.e. job_title, job_description, location_city, location_region, date_posted, and enterprise_name). A common problem was the job title field containing extraneous information (e.g. ".NET Developer, Stoke-on –Trent, £35-40K"), which needed removing. The data are then cleaned using Regex functions. Some basic language processing was done on the job_description field (e.g. removal of stop words).

Duplication may occur both within a job portal as well as between them. It was therefore decided to implement a 2-stage approach by first identifying duplicates within each portal with a second stage to identify duplicates between them. The reasoning was that patterns of duplication may be specific to each job portal, in which case better results would be achieved if the algorithms were tailored to each portal.

The first Dedupe step involves an initial run which uses logistic regression to produce a similarity score based on the variables selected for matching. This is followed the active learning step where potential duplicate advertisements are displayed on the screen and a clerical decision is made on whether or not they are the same vacancy, with possible responses being "yes", "no" or "uncertain". These training examples are skewed towards those where there is some uncertainty. This process is then repeated a number of times to create a training data set. This is used to re-weight the logistic regression algorithm, which is rerun against the input file. This produces a dataset which clusters those records deemed to be duplicates providing a mechanism for duplicate records to be removed.

Although this did identify a considerable number of genuine duplicates, inspection of the results showed that a lot more work would be needed to achieve high quality results:

- More comprehensive strategies are needed for cleaning and standardising data.
- More and better quality training data is needed. In particular, clearer principles are needed about how to decide whether two advertisements are likely to refer to the same vacancy.
- More thought is needed on how to extract features from the job description field (which contains a lot of unstructured text).
- The complete job description field should be captured and processed as opposed to a 'snippet'.

In summary, this supervised machine learning approach seems to be viable, but more work is needed to produce high quality results. Producing the required volumes of training data may not be practical as this may involve a lot of clerical resource that has not been factored into the pilot.

6.6.3 Matching

The UK has also undertaken a small scale experiment into matching advertising businesses on job portals with enterprise units on the job vacancy survey. This focused on the approximately 1300 largest employers that are always in the survey. The initial step was to explore a range of data transformation[9] and automated matching methods[10] and applying them in sequence to produce an overall set of matched results.

Several problems were identified with this approach:

- The advertising business on a job portal may use a trading name that is different from the enterprise name recorded in the survey data.
- For enterprises part of a complex enterprise group, the advertising business may be different from the reporting unit.

An investigation was made into the possibility of matching the advertising business with the trading name from UK Companies House[11] data to get the enterprise name. However, this uncovered another problem with duplicate business names. Further work was done looking into the feasibility of using location information, but this proved to be very complicated.

In combination these approaches were able to match job portal data for just over 27 per cent of these large enterprise units and so there is a long way to go, even for this small subset. Further improvements may be possible through a supervised machine learning approach. There may also be information available from the business register that could help resolve these some of matching issues. However, it may be that high quality results on a larger scale will require a significant investment in manual matching and resolution.

6.4.4 Next Steps

The main focus now is on continuing to develop methods for matching job portal and survey data to reconcile and measure the gap between job portal coverage and the survey. An interesting recent discovery is that the Indeed job portal maintains a set of structured web pages for large enterprises which includes a list of current job vacancies and an overall count[12]. An initial manual check found that for the most part, this job vacancy count matched exactly, or was very close to, the number of vacancies advertised on the enterprise's own website. Therefore, it appears that Indeed already scrapes and compiles data directly from many enterprise websites. This may be a fruitful avenue as this may provide access to data fairly easily from a definitive source that is not subject to the usual problems of duplication. Also, Indeed operates in over 50 countries so similar data may be available for other countries in the ESS.

[9] These were: simple match, lower case, words ordered alphabetically, removal of non-distinguishing strings (e.g. "Ltd"), removal of multiple white spaces

[10] These were: Levenstein distance, N-grams, Jaro. A match rating/Soundex approach was also applied, but this had no effect and was removed.

[11] UK Register of companies

[12] http://www.indeed.co.uk/cmp/Tesco/jobs?clearPrefilter=1#cmp-menu-container

Annex: Description of Current Job Vacancy Survey/Statistics

1. Germany

The Germany system of statistics on job vacancies consists of two main pillars: The job vacancy survey and the survey on vacant posts registered at the Federal Employment Agency (FEA). Furthermore, the FEA runs and online job portal and job robot that web scrapes vacant posts at enterprises, which are however used for statistical purposes to a very limited extent.

The German *Job Vacancy Survey* is carried out by the Institute for Employment Research (IAB – Institut für Arbeitsmarkt- und Berufsforschung). The IAB, which is based in Nuremberg, was set up in 1967 as a research unit of the former Federal Employment Service (Bundesanstalt für Arbeit) and has been a special office of the Federal Employment Agency (Bundesagentur für Arbeit/BA) since 2004.

The IAB Job Vacancy Survey is a representative survey including all economic sectors and establishment sizes in Western and Eastern Germany. The regular surveys of a representative selection of establishments and public institutions are geared towards personnel representatives and/or business managers with personnel responsibility. The survey started with a written questionnaire in West Germany in 1989 and has since been repeated every year – always in the fourth quarter – as cross-sectional survey. In 1992 this was extended to include former East Germany. In 2006, the collection was expanded to a quarterly collection to incorporate EU regulation No. 453/2008. This involved supplementing the written questionnaires of the fourth quarter IAB Job Vacancy Survey with short follow-up telephone interviews in the following three quarters. The survey is carried out by economic research institute Economix Research & Consulting, located in Munich.

The survey run in the fourth quarter is divided into four sections:

a) Main questionnaire: number and structure of jobs and of vacancies, newly recruited staff members, cancelled recruitment processes.

b) Special questionnaire: current labour market policy topics (e.g. recruiting decision during the economic crisis, recruitment opportunities of long-term unemployed, labour market policies)

c) Last case of successfully hiring a new employee in the past twelve months

d) Last case of terminating the recruitment process in the past twelve months

The quarterly follow-up surveys in the first, second and third quarter following the first wave interview include only a smaller number of variables, focussing on the variables required by the EU regulation.

The population used for the sampling originates from the currently available address stock of the Federal Employment Agency's (BA) register of employees. Usually this address stock is about eight months old at the time of sampling. This includes all establishments with at least one employee subject to social insurance contributions. As the labour markets in Western and Eastern Germany differ significantly, random samples are drawn separately for the two regions, stratified by 23 economic activities and seven establishment size classes (number of employees subject to social security contributions).

The gross sample of the first wave interview in 2015 included about 75,000 enterprises. As about 20% of the enterprises take part in the survey, the net sample was about 15,000 enterprises. The interviews of waves two to four are conducted as a sub-sample of the first wave participants where the net sample size is about 9,000 enterprises (for details. see Mozcall et al. 2014; Brenzel at al. 2016).

The weighting factors of the job vacancy survey are computed using a generalised regression estimator (GREG) that includes benchmarks regarding the number of businesses and employees subject to social insurance contributions for each stratum. It equally includes a correction for non response. which uses further auxiliary variables that were identified in a detailed non-response analysis (see Brenzel et al 2016).

The other major data source is the *statistics of registered job vacancies*, which is a set of register based statistics including vacancies for which the employers have mandated the Federal Employment Agency to provide placement services. It is implemented by the FEA on the basis of the administrative documents. This statistics only cover a subset of the vacancies (about 70% of the vacancies measured by the job vacancy survey), but as it is based on the register, allows for much more detailed breakdowns. The variables available in the statistics of registered job vacancies include the place of work, the occupation required, whether the job is subject to social insurance contributions, the type of contract (open-ended vs. temporary), full-time or part-time work as well as the economic activity. Furthermore, the data can be used to analyse the duration of vacancies (see Bundesagentur für Arbeit 2016).

In addition to the data from the statistics on registered job vacancies, the FEA disposes of further valuable data, which are currently only exploited for statistical purposes to a limited degree. The *job portal* of the FEA, in addition to the vacancies for which there is a placement mandate, contains further vacancies that are supplied by other cooperating job portals or directly by employers. In total, the FEA job portal includes about 1.2 million job advertisements. In addition, the FEA has commissioned the development of a *job robot* that web scrapes the web sites of employers and serves as an additional input for the FEA's placement services. The job robot currently includes 780,000 job advertisements. Both the job portal and the job robot are including duplicates and are currently only used statistically (together with the statistics on registered vacancies) for the calculation of a job vacancy index (BA-X).

2. Greece

Target population: The statistical population is all the enterprises employing at least one (1) employee and belonging to Sections B-S of NACE Rev.2.

Reference area: Greece, total

Reference period: The reference period of the data on Job Vacancies is one calendar quarter

Periodicity: The results of the Job Vacancy Survey are available 70 days after the end of the reference period.

The Job Vacancy Survey is conducted on a sample designed from the Statistical Business Register of ELSTAT. More specifically, for every two-digit code of economic activity a number of enterprises are selected for each one of the seven size classes, in which the enterprises are classified on the basis of their annual average employment. Until the 4th quarter of 2015, the survey was conducted based on a sample of 6,774 enterprises and services for all quarters. From the 1st quarter 2016 onwards, the survey was redesigned and it is being conducted on a sample of 7,511 enterprises and services. The data are collected through paper questionnaire. Non-response is addressed through telephone contacts with the enterprises, reminders sent by fax or e-mail or personal visits to the enterprises.

The concepts and definitions of the basic variables used for the compilation of Job Vacancy Statistics are laid down in the European Regulations (EC) No 453/2008, (EC) No 1062/2008 and (EC) No 19/2009.

A vacant post which is going to be filled by any of the following cases is not considered as Job Vacancy:

- An apprentice without remuneration coming either by the employer or through the Social Security Funds
- Contractors which are not on the payroll list,
- Personnel that are re-hired or return to the enterprise after a holiday paid
- Internal movement of a member of personnel inside the enterprise

Job Vacancies to be filled in immediately are job vacancies for full or part-time employment which are to be filled in within a period **not** longer than three months (starting day of the quarter is considered the first day of the third month of every calendar quarter).

Job Vacancies in the near future are job vacancies for full or part-time employment which are to be filled in within a period longer than three months (starting day of the quarter is considered the first day of the third month of every calendar quarter).

Full-time Job Vacancies are posts which are to be filled by employees whose regular working hours are the same as the collectively agreed or customary hours worked in the enterprise, even if their contract is for less than one year.

Part-time Job Vacancies are posts which are to be filled by employees whose regular working hours are less than the collectively agreed (set out in the collective or industry employment agreement) or customary hours worked in the enterprise.

Collected variables:

- Number of Job vacancies
- Number of occupied jobs.

Estimates and domains:

- Totals of Job vacancies
- Totals of occupied jobs adjusted from LFS data

broken down by:

- 1-digit Section NACE Rev.2
- size of enterprise
- 1-digit level of occupation groups ISCO-08

Main users and usage: The main users of the Job Vacancies Survey are international organizations (Eurostat), as well as many national authorities, services and institutions (Government, Banks, Universities, Research Institutes, etc.). Furthermore, among the users are the press, researchers and the general public.

Other issues: ELSTAT is also exploring the possibility to collect data from administrative sources (Social Insurance Institute –IKA, “ERGANI” project), in order to enhance the quality of data and reduce the administrative burden of enterprises.

3. Slovenia

The observed population are all business entities as a whole (LU), registered on the territory of the Republic of Slovenia which had at least one employed person when the sample was prepared. Natural persons who have no employees besides themselves are not included within the target population. Included are business entities with registered main economic activity from B to S.

The reference period is the last working day in the middle month in every quarter.

Definition: job vacancy (JV) is defined as a post (which has been newly created. is unoccupied or will shortly become free) for which the employer is actively seeking a suitable candidate outside the enterprise and which will be filled immediately or in the near future.

Job vacancies do not include posts that will be filled by unpaid trainees, contract workers (who are not on the payroll), persons returning from paid or unpaid leave, or persons who are already employed in the firm and who will occupy a post as a result of the reorganisation of the firm.

Definition: occupied post (OP) is a post filled by a person in paid employment who has compulsory pension and health insurance on the basis of an employment contract or who is in an employment relationship. The employment relationship may be established for a fixed or indefinite period on the basis of full-time or part-time work. In the number of occupied posts are included: persons employed by legal or natural persons and posted workers since July 2009 (the workers who are sent abroad to work or study; they get wages from Slovenian employer). But in the number of occupied posts are not included persons who are recipients of parental compensation (persons on maternity leave-since January 2009) and persons who are on long-term sick leave more than 30 working days (since

January 2013). Long-term sick leaves are no longer covered by the employer but by the Health Insurance Institute of Slovenia. In such cases, it is likely that employers at the same post hire another person as a replacement. This has improved the quality of data and reduces the likelihood of double-counting the number of occupied posts.

Statistics:

- number of job vacancies
- number of occupied posts
- the job vacancy rate

All three indicators must be broken down by the sectors of activity and by the size of business entity:

- 2 size classes: total and business entities with 10+ employees
- 18 industry sectors

All three indicators in data dissemination are presented as:

- Total population.
- Total population broken down by activities.
- Business entities with 10+ employees.
- business entities with 10+ employees broken down by activities

Methods of data collection:

In 2015, SURS began to collect the job vacancy data independently with a sample survey. The collection methods are as follows:

- by e-STAT application (WEB) – first 14 days after the reference day.
- by CATI – next 14 days after collecting the data via WEB.
- by the Contact Center at SURS. the reporting units contact them by phone or by email – the whole month after the reference day.
- by overtaken the data collected by Employment Service of Slovenia (for employers of the public sector and for state owned companies. They are still obliged to report job vacancies to Employment Service of Slovenia).

Finally we combine the collected data with sample survey and the administrative data to calculate the estimates of the number of job vacancies. To calculate the number of occupied posts is using the data overtaken from Statistical Register of Employment and it also refer to the last day of the middle month in the quarter.

Current practice at SURS:

The sample is prepared in the beginning of the year and is valid for the calendar year (that means that new establishes enterprises will not normally be included in the frame). The sample includes all business entities which had at least one employed person when the sample was prepared (the reference month for persons in employment is October of the previous year). The employers of the public sector and for state owned companies are not included in sample.

At the end of 2015 in the sampling frame for 2016 were a little more than 61,500 businesses which had at least one person in employment. From the sampling frame around 8,900 business entities were selected on the basis of random selection, which is 14.5% of population. The number of employers of the public sector and for state owned companies is around 3,400. Data for this part of the population are still taken over from the administrative source. The final sample size is about 12,300 business entities (or 20% of population).

Publishing:

Results in the First Release are published as absolute data on the number of vacancies and occupied posts by sector of activity and by size of the business entity. Besides the absolute data the vacancy rates are also published. We publish the original and the seasonally adjusted data and for seasonal adjustment of time series we use **JDemetra+** software, i.e. the TRAMO/SEATS method.

4. Sweden

The Swedish Vacancy Survey is a sample survey carried out by Statistics Sweden on a quarterly basis. The results are published about six weeks after the end of a quarter.

The target population are all active businesses and organizations in the public and private sector (economic sector A-S) with at least one employee. The Swedish Business Register is used as frame. The population is stratified on industry and size of the establishment, measured by the number of employees. Strata for sizes above 100 employees are completely enumerated, for the remaining strata a sample is selected. The sampled units are the legal units for the public sector and the local units for the private sector.

A total of about 17,350 units are included in the survey (16,700 from the private sector, 700 from the public sector). Units in completely enumerated strata report for each month in the quarter. The sample is divided in three parts and each unit reports for one month in the quarter. For the whole sample, the reference period is a Wednesday in the middle of the month.

Data is collected primarily by a web questionnaire, but a paper questionnaire is also available.

The sample is rotated in order to decrease response burden, a sampled unit remains in the sample for a maximum of five years.

In order to get a fair picture of the survey, it is important to understand the definitions of the concepts that the survey is trying to measure. Any new source, such as data from job portals or web pages, must be able to directly measure, or support measurement of, similar concepts in order to be useful.

A *Job opening* means that the employer has begun recruitment but has not yet filled the position. A job opening can be *occupied* or *unoccupied*, where unoccupied implies that the position can be filled immediately. A *vacancy* is an unoccupied job opening. *Recruitment* refers to the availability of a vacant job that is accessible for external applicants.

The survey collects the following variables:

- Number of job openings on the reference day (for each unit in the survey)
- Number of vacancies on the reference day (from each unit in the private sector)
- Number of employees (from register)
- Number of new recruitments (from short-term employment)

The following estimates are obtained:

- Totals of job openings
- Totals of vacancies
- Recruitment rate
- Vacancy rate

Recruitment rate is the number of job openings divided by the number of employees in the reporting group. The *vacancy rate* is the number of vacancies divided by the number of employees in the reporting group. The *average recruiting time* is the number of job openings divided by the number of new recruitments in the reporting group.

These estimates are broken down by sector and industry, size of enterprise, and region (NUTS 2).

The main users of the figures are the Swedish Unemployment Agency, the Swedish National Mediation Office, other Government offices, the National Institute of Economic Research, the Swedish Central Bank, universities and researchers.

5. United Kingdom

The Vacancy Survey is a statutory monthly survey of businesses. The survey asks a single question: how many job vacancies did a business have in total (on a specified date) for which they were actively seeking recruits from outside their organisation. Results from the survey cover all sectors of the UK economy and all industries with the exception of employment agencies and agriculture, forestry and fishing.

The Inter-Departmental Business Register (IDBR) is used as the sampling frame. The survey is sent to approximately 6000 business every month with approximately 1,300 large businesses included every month and the remaining 4,700 smaller businesses sampled randomly on a quarterly basis. The headline series are based on 3 month moving averages by industry type and enterprise size (by employment count).

References:

Bundesagentur für Arbeit. 2016: Statistik der gemeldeten Arbeitsstellen. Qualitätsbericht. Nuremberg. (in German only) Available at: https://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Qualitaetsberichte/Generische-Publikationen/Qualitaetsbericht-Statistik-gemeldete-Arbeitsstellen.pdf (accessed 1 November 2016)

Kettner. A. and Vogler-Ludwig. K.. 2010 “The German Job Vacancy Survey: An Overview” in “1st and 2nd International Workshops on Methodologies for Job Vacancy Statistics. Proceedings”. Eurostat (Accessed 24 Oct 2016: http://ec.europa.eu/eurostat/documents/3888793/5847769/KS-RA-10-027-EN.PDF/87d9c80c-f774-4659-87b4-ca76fcd5884d)

Körner. T.. M. Rengers et al.. 2016. “Inventory and qualitative assessment of job portals” Deliverable 1.1. Work Package 1 Web scraping / Job vacancies of the ESSnet on Big Data. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_draft_v5.docx (accessed 1 November 2016)

Moczall. A. et al.. 2014. “The IAB Job Vacancy Survey. Establishment Survey On Job Vacancies and Recruitment Processes Waves 2000 to 2013 and subsequent quarters from 2006” Nuremberg. Available at http://doku.iab.de/fdz/reporte/2015/DR_04-15_EN.pdf (accessed 1 November 2016)

Brenzel. H. et al.. 2016 “ Revision of the IAB Job Vacancy Survey. Backgrounds. methods and results ” Nuremburg. Available at http://doku.iab.de/forschungsbericht/2016/fb0416_en.pdf (accessed 1 November 2016)