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 1.3

Final Technical Report (SGA-1)

Version 2017-07-08

Prepared by:

Nigel Swier, Frantisek Hajnovic (ONS, UK)
Ingegerd Jansson, Dan Wu (SCB, Sweden)
Boro Nikic (SURS, Slovenia)
Christina Pierrakou (ELSTAT, Greece)
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

Table of contents:

Executive Summary:

A team of six National Statistics Institutes (NSIs) within the ESS is investigating the feasibility of using job advertisement data scraped from the Internet to improve official estimates of job vacancy statistics. This work is being taken forward as part of a Big Data ESSNet funded by Eurostat. The pilot began in February 2016 and finishes in May 2018. This report is a summary of progress from the first stage of the pilot.

The production of job vacancy statistics is subject to EU regulation 453/2008, which sets out certain minimum requirements for data collection including estimates of the stock of current job vacancies produced each quarter and broken down by industry sector and enterprise size. However, these statistics have no information about the types of jobs, or their location. On-line job advertisements contain text rich information that could enhance the utility of the official statistics. In addition, job vacancies are a leading economic indicator and so real-time availability could be used to better inform economic policy.

The on-line job vacancy “landscape” varies considerably between EU member states in terms the importance of on-line advertising channels, the number of on-line job portals and the number of jobs advertised on-line. Also, while job vacancy statistics are subject to regulation, the specific arrangements vary between countries. Therefore, the approach to be taken by each NSI needs to reflect the specific situation within each country.

Job portals fall in three categories:

i. job boards (where employers place ads)
ii. job search engines (which crawl job boards and republish ads)
iii. hybrid portals (that do both)

In addition, many employers also advertise vacancies on their own website. This means that a single job vacancy is typically advertised multiple times and so combining data from multiple on-line sources leads to issues with duplication. There are other issues in aligning the target concept of a job vacancy with the target measure including, ads with multiple vacancies and ads for non-existent vacancies. In addition, a vacancy may be created before it is advertised on-line and may persist for some time after the closing date of the advertisement.

On-line job advertisements usually consist of a small number of structured elements and a larger amount of text containing the full job description. However, even the stuctured elements can be messy and a lot of effort is required to clean and classify data prior to analysis. In addition, job portals usually have their own taxonomies for classifying data. There are also important legal and ethical issues with regards to web scraping to be aware of, particularly when scraping large volumes of data. Therefore, transforming web scraped data into data ready for analysis is a resource intensive activity.

For these reasons, NSIs (particularly in larger countries) should be aiming to build partnerships with job portal owners, or others with access to job vacancy data, rather than looking to build their own large scale web scraping systems. One such opportunity is with the EU Centre for Vocational training (CEDEFOP), which is currently undertaking an EU-wide project to web scrape job vacancy data for all

EU member states. This pilot is building a partnership with CEDEFOP with a view to coordinating activities. In the longer term these data may become available for NSIs within the ESS.

Although the level of on-line coverage of job vacancies varies between countries, vacancies advertised on-line are invariably a subset of vacancies advertised through all channels. Therefore understanding this gap is the key to understanding how these data could be used alongside the official estimates. There are two main approaches to trying to measure this gap, to understand this dimension of quality and to develop strategies for dealing it:

i. Collect data on advertising channels via a survey
ii. Link web scraped counts by enterprise and linking it to data from the job vacancy survey.

The potential of linking micro data from the job vacancy survey data to better understand the on-line data is a promising avenue of research as it enables this gap to be understood at the level of individual enterprises. However, this approach comes with its own problems, including the difficulties of matching on business name between sources. Also many on-line jobs are advertised through employment agencies and so direct matching with the survey is not always possible.

All these issues mean that this work package has not yet managed to produce much in the way of concrete estimates, even on an experimental basis. However, it is expected that some experimental results will be produced by the end of the ESSNet. Longer term it is very unlikely that on-line job ad data would ever completely replace the existing job vacancy survey. A survey would continue to be needed to provide a benchmark, with the on-line data providing the additional detail.

1. Introduction

This report is a summary of the work completed by the Big Data ESSNet WP1 – Web Scraping for Job Vacancy Statistics during SGA-1. It describes the work completed and aims to bring to together the key lessons learnt since the start of the pilot in February 2016. This is the third deliverable in this work package, the first two being:

i. Qualitative Assessment of Job Portals (delivered July 2016)[1]
ii. Interim Technical Report (delivered December 2016)[2]

There are six countries participating in this pilot:

- Germany (DESTATIS)
- Greece (ELSTAT)
- Italy (ISTAT)
- Slovenia (SURS)
- Sweden (Statistics Sweden)
- United Kingdom (ONS)

Italy's involvement is limited to collaborating on the use of the methods developed by a separate ESSNet work package (WP2 - Web scraping for enterprise statistics), led by ISTAT. The primary focus during SGA-1 has been on the web scraping of job portals and SGA-2 will focus more on the potential of scraping job vacancies from enterprise websites.

There has also been some collaboration between WP1 and WP2 on legal issues on web scraping, which was published as a WP2 deliverable[3]. The report summarizes the relevant legislation at EU level and the current situation in the six countries participating in WP2. All countrie have national statistical laws and copyright protection, and the main differences between countries seem to be in interpreting the laws regarding web scraping. The report gives some recommendations for web scraping by statistical agencies, such as being transparent, have a clear policy, communicate through web sites, and cooperate with web site owners. Further, it is suggested that common ethical guidelines for web scraping are deployed throughout the EU.

Although on-line job portals exist in all countries, the job market in each country and in particular the numbers of job vacancies varies considerably (Table 1). Unsurprisingly, the number of job vacancies in each country is broadly in proportion to the size of the workforce. However, Greece has a much lower number of job vacancies as a proportion of total employment and very high unemployment, and so economic factors are also important. There are structural differences between economies that are known to be related to how jobs are advertised. For example, Germany has a high number of large enterprises (over 250 employees) and such businesses a more likely to advertise using on-line job portals[4]. Similarly, there are differences between countries in the proportion of the

[1] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_final.docx
[2] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/64/WP1_Deliverable_1_2_final.pdf
[3] Stateva e al (2016), "Legal aspects related to web scraping enterprise websites" (Section 4 p.17) ;
[4] Studie von index.de siehe Mail vom 21.10.2016 vgl. S. 14, S.23

workforce in self employment, which will also influence the overall demand for employed labour in each country.

Table 1: Selected labour market and enterprise statistics by country (2014 Q4[1])

Country	**Job Vacancies** (thousands)	**Total Employment 15-64** (thousands)	**Job vacancy rate** (%)	**Unemployment 15-64 (%)**	**Enterprises > 250 employees** 2013	**self - employed 15-64** (thousands)	**self- employed 15-64** (% of total employed)
Germany	1,056	40,770	2.6	3.8	10,361	3,724	9.1
Greece	19	3,585	0.3	23.7	376	1,059	29.5
Slovenia	15	912	1.7	8.1	200	104	11.4
Sweden	126	4,742	2.0	6.6	961	418	8.8
United Kingdom	742	30,619	2.5	4.7	5871	4330	14.1

Sources: Eurostat, OECD

[1] except for enterprises with more than 250 employees where reference year is 2013

Furthermore, there also seems to be relationship between the overall size of the on-line job market and the complexity of the on-line job vacancy landscape in different countries. For example, there are over 1600 job portals in Germany with many jobs advertised multiple times across many portals. In contrast, Slovenia less than 30 job portals, most of which are very small. Therefore, the complexities faced between large and small countries are very different.

There are also operational factors that constrain the approach that each NSI is able to take. At Statistics Sweden, a lack of clarity legal issues around web scraping means that it has not been possible to perform any web scraping experiments. As a consequence, efforts have focused on obtaining data directly from the Government Employment Agency and large job portals. In the case of DESTATIS, the job vacancy survey (JVS) is the responsibility of another public institution and so the matching and linking approaches undertaken by other countries has not been possible. Instead, comparisons have been made with the job vacancy survey using aggregate data.

For all these reasons, each NSI has pursued a different path but information is shared and collaboration takes place where possible. This approach has had provided an opportunity to test out different ideas and identify those that show the most promise. This report pulls together and summarises the main themes running through each of the pilots, following the broad structure agreed for each Big Data ESSNet as a whole, namely:

1. Data access
2. Data handling
3. Methodology
4. Statistical outputs
5. Future prespectives

The details of each country pilot are presented as separate annexes.

2. Data Access

The approaches to accessing on-line job advertisement data can be divided into two broad types: direct web scraping and arranged access.

2. 1 Direct web scraping:

This involves using web scraping techniques, or a public API, to collect data from on-line sources without an explicit data access agreement from the website owner. Target websites may include either job portals or enterprise websites. The main advantage of direct web scraping is that data can be captured and assessed quickly. This may be achieved either through simple "point and click" web scraping tools or more programmatic approaches. Direct web scraping also offers a high degree of control over what is collected and in a production scenario could provide an opportunity to produce outputs in near real-time.

The main disadvantage of direct web scraping is that many websites have restrictions on what content may be scraped from a web site or may restrict it entirely. As discussed in the WP2 Legal Report, the specific legal risk is around "sui generis" database rights, which are a form of ownership right pertaining to data that apply when scraping "substantial parts" of a website (Stateva et al, 2016, p.4). However, NSIs should also consider the ethical and reputational risks of web scraping. These risks can be managed by following principles of web scraping "netiquette", such as respecting the robots.txt exclusion protocol (ibid, p.21).

There are also other practical issues for NSIs around web scraping. Some NSIs may have restrictions on the use of web scraping tools on corporate IT environments. Programmatic approaches also require good coding skills and detailed knowledge of website structures. Developing production scale web scraping applications with repeated collections may require a substantial effort to both in terms of development and ongoing maintenance. For these reasons, arranged access with a job portal website (or other data owners) may be a preferable approach for many NSIs. However, in some cases, direct web scraping may often be necessary. For example, scraping job ads from enterprise websites sampled from the business register would probably need to be scraped directly by the NSI.

2.2 Arranged access:

This involves an explicitly agreed data access arrangement with either the website owner, or other organisations holding these data including job portals, employment agencies (government and private) and data aggregators. There are several international companies that scrape job advertisement data from the web, process it and use it provide data and analytical services (e.g. Textkernel, Burning Glass). Data may be supplied as files, through an on-line tool, or via an API.

There are a number of advantages of accessing data through an explicit agreement from job portal owners rather than web scraping the data. Most importantly, it removes any uncertainty over legal issues in accessing data and can provide assurances that any outputs can be published. An explicit agreement may also offer a route for accessing historical data. Finally, the job portal owner may also be able to provide insight into the methods used to collect and process the data.

The main disadvantage of pursuing direct access arrangements is that developing these relationships can be time consuming and need to be managed carefully. The UK experience is that commercial companies, including job portals, expect payment in return for any services provided, although this seems to be less of an issue in other countries. The UK response has been to explore partnership models, which could offer in-kind services, such as linking their data with the JVS, in return for data access. However, this approach needs to follow rules on Government procurement and so this has required going to market to ensure that the same opportunities are available to all potential data suppliers.

Sweden managed early on to secure access to a large amount of data from its government employment agency while Slovenia has gained access to advertisements for all public sector jobs which must be advertised through the Slovenian government employment agency website. DESTATIS are in discussions with the Federal Employment Agency about obtaining access to its job portal database, the largest job portal in Germany.

In 2015, the EU Centre for Vocational Training (CEDEFOP) funded a web scraping pilot of selected job portals in five member states. While analysis of the UK results showed a fairly large gap between the number of scraped job advertisements and official job vacancy estimates, the prospect of a longer-term collaboration seems appealing. In early 2017, CEDEFOP launched the second phase which plans to develop a system for collecting on-line job advertisement data for all EU member states. It is becoming clear that different parts of the European Commission and other public bodies have an interest in this kind of data. There is also a growing recognition that it is not efficient in the long-term to have duplicate systems collecting and processing the same kind of data. A formal agreement is in now place between CEDEFOP and Eurostat which will enable to WP1 to collaborate with CEDEFOP and to ensure that statistical requirements are taken into account. Expertise in the areas of quality and statistical measurement means that the ESS can play an important role in the appropriate use of these data for policy making purposes. This CEDEFOP-led project could be the long-term solution for ESS to be able to access this kind of data.

2.3 Approaches to JV data access by country

All countries have managed to gain some form of access to job portal data, either by web scraping directly, or in the case of Sweden, through agreed access via an API. Although the initial plan was to explore web scraping enterprise websites during SGA-2, Slovenia and the UK have already started. The UK decided to do this in order to get an earlier insight into the relative performance between enterprise websites and job portals compared with the job vacancy survey. Most countries have data access agreements with their Government Employment agency or taking steps towards this.

The various avenues to data access by each country are summarised in Table 2.

Table 2: Investigation of On-line Job Vacancy Sources by Country

Country	Direct Web scraping		Agreed Access			
	Enterprise websites	Job Portals		Government Employment Agency	Private Employment Agencies	Data aggregators
Germany		Yes		Active discussions		
Greece		Yes		Considering		
Slovenia	Yes	Yes	Yes	Yes	Yes	
Sweden			Yes	Yes		
United Kingdom	Yes	Yes	In progress			CEDEFOP

3. Data Handling

3.1 Web scraping:

Several countries have experimented with point and click web scraping tools (e.g. Import.io, Content Grabber) for obtaining data. Although these experiments have generally been small scale one-off data collections, Slovenia have used this successfully for an experimental weekly collection system. Over the last six months the UK have followed a programmatic approach to web scraping developing a framework using Python Scrapy. This involves scraping daily counts from a sample of companies from a selection of job portals and also from the enterprise websites of those same companies. Other approaches include the use of job portal APIs, which is an easier and more reliable way of obtaining these kinds of data.

3.2 Data transfer and storage:

For the most part, the volumes of data that have handled so far in this pilot are not very large and so big data solutions have not been necessary. However, the UK pilot is using NoSQL data storage (i.e. Mongo DB) partly because a similar approach has been established for other web scraping projects, but also because a schema-less NoSQL database is flexible for scaling up, either in terms of expanding the list of target companies, scraping new portals or collecting new variables. The UK web scrapers are running on a Google compute platform. Sweden use Python for data handling and SQL server for data storage. Analysis tools used within the pilots are a mix of Python libraries (e.g. pandas, maplotlib) and Excel.

3.3 Data cleaning and de-duplication:

The raw data obtained from a job advertisement generally requires a lot of processing before it can be analysed. For example, job title fields often contain extraneous information, such as job location, key skills, and salary. This is because employers try to attract potential job seekers by stacking the job title field with other key information like skills required and salary.

Duplication of job ads is a fundamental quality issue with multiple job portals. It can also be a problem within portals, particularly for job search engines that pick up job ads from other portals. Job search engines may take steps to remove duplicates but the effectiveness of these procedures is often variable. Duplication methods were explored as part of a "virtual sprint" held on 28-29 July[5]. Essentially, these methods involve matching common fields, comparing text content and then calculating a similarity metric to establish the likelihood that two job advertisements are the same. The Slovenian pilot has a duplication issue with their weekly collection where jobs advertisements will typically span more than one reference period. However, these duplicates can be identified quite easily as any job ad that had the same URL as an existing record.

To remove duplicate adverts that do not have a unique URL, the first step is to prepare and standardize the data fields that are common and that can be compared to all data sources (i.e. job title, location, company name, date posted, job description). This involves text normalisation procedures such as, removal of white spaces, case standardisation and removal of stop words or other extraneous text, typically using regular expressions (regex).

The next step involves calculating a similarity metric to identify any likely duplicate job ads. One approach explored by the UK involved using Python Dedupe, which is designed to identify duplicate records using supervised learning methods. This uses an initial match using logistic regression and then identifies marginal cases for clerical resolution. The decisions of this clerical process are then reincorporated into the machine learning algorithm, to be applied for automated removal of duplicates. Other matching methods were explored by Sweden and included, Levenshtein distance and longest common substring distance, although Jaccard similarity performed best.

The initial focus was on the structured data fields rather than the unstructured content of the full job description. This was mainly because this information is often difficult to scrape from websites in full and often only a "snippet" of a certain number of characters is readily available. This information may often be needed to achieve a good quality de-duplicated data set, especially with a large number of records.

3.4 Classification and data enrichment

Once data fields have been cleaned, derived variables can be added to enrich the data for analysis purposes. This is particularly important for deriving occupation codes from job title and standard geographies from job location fields. However, to date this is not an area that we have had time to explore in any detail.

[5] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP1_2016_07_28-29_Virtual_Sprint_Notes. These are also documented within the WP1 - interim technical report

3.5 Conclusion:

Our consensus is that the effort required to scrape and transform large amounts of raw data from job portals into a form ready for analysis, is considerable and may not be the best use of scarce resources, especially where there are vast amounts of data. It is probably better for NSIs to focus on obtaining data that has already been prepared for analysis and/or else focus on other aspects where NSIs can add more value. In particular, matching on-line data on job ads with survey data seems to be a good example of where NSIs would be able to add value. It could also be the basis on which partnerships with data owners are established.

4. Methodology

So far, the main methodological focus of this work package has been on data quality and matching. It is clear that there are fundamental issues around the quality of job advertisement data, what these represent, and how they compare with estimates from official surveys. Critically, not all job adverts are advertised on-line and there is geographic variation both in the types of jobs advertised on-line and the types of jobs (Carnvale, 2014). The ultimate goal would be to develop methods for combining on-line job ad data with the existing job vacancy survey so as to produce statistics that were both more granular and fully consistent.

There are alternative approaches that could be seen as intermediate steps towards this goal. One could be to focus only on producing statistics of on-line job vacancies only. However, even this is a major challenge, not only because of the many different sources of on-line data and the problems of duplicated advertisements but also because of other measurement issues such as multiple jobs per advertisement, and so called 'ghost' vacancies (i.e. job advertisements for a non-existent vacancy)[6]. Another approach could be to focus more on developing methods for using on-line sources to measure change over time rather than to estimate levels. For example, there is the need to account for the variable growth and decline of different portals over time. For these reasons, the production of meaningful statistical outputs requires data of suitable quality.

4.1 Definitions

There are differences between the target concept of official surveys and what can be practically measured from on-line job advertisements. 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

[6] These issues are discussed in detail as part of the Interim technical report.

(b) which the employer intends to fill either immediately or within a specific period of time." [7]

EC regulation 453/2008 has several mandatory elements:

- Quarterly data that has been seasonally adjusted
- Data broken down economic activity (using NACE[8])
- Data is relevant and complete, accurate and comprehensive, timely, coherent, comparable, and readily accessible to users.

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).

The official definition of a job vacancy does not correspond exactly to what is indicated by an advertisement. A vacancy must be created before it is advertised and will normally exist for a period after the closing date of the advertisement. Therefore the stock of current vacancies as measured by the job vacancy survey should be somewhat larger than the corresponding stock of live job advertisements for those vacancies. Research from Slovenia suggests that there is on average a lag of 45 days from when a job is first advertised to when it is actually filled. These differences need to be taken into account when directly comparing data sources.

4.2 Quality Assessment Frameworks

A third WP1 "virtual sprint" was held on 1st of February 2017, focusing on quality. We decided to explore two different quality frameworks:

1. The Quality Assessment Framework used by Statistics New Zealand as reporting tool for administrative data quality. The aim was to test the suitability of this framework for web-scraping for on-line job advertisements. This was in part in response from some initial proposals put forward by WP8 for approaching big data quality

[7] 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

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

2. The Framework for the Quality of Big Data Framework for the Quality of Big Data developed by the UNECE Big Data Quality Task team.

Sweden, Germany, Greece and the United Kingdom focused on the first while Slovenia worked on the UNECE framework. A report on this virtual sprint is available on the ESSNet wiki[9] but a summary is provided below.

4.2.1 Statistics New Zealand Quality Framework for Administrative Data:

The framework is designed to support the measurement of total "survey" error (Reid et al, 2017). The framework comprises of two parts. Phase 1 applies to a single dataset in isolation and considers errors in terms of measurement (variables) and representation (objects). The measurement side describes the steps from the abstract target concept through to an edited value for a concrete variable. The representation side describes the definition and creation of the elements of the population being measured, or 'objects'. The output is a single source micro dataset that could be used for different statistical purposes.

Phase 2 is used when producing and output for a particular statistical purpose, and may involve combining data from several different sources. In this phase, the measurement side focuses on relevance error, mapping error between sources and comparability of adjusted measures. The representation side focuses on coverage error of linked datasets, identification error (the alignment between base units and composite linked units) and unit error, which may be relevant if the output involves the creation of new statistical units.

Issues were identified for all error types in Phase 1 with quite a number of issues identified under the categories of validity and frame errors. The assessment of Phase 2 was more complicated and was not completed. One reason for this could be that the pilot has not yet undertaken comprehensive work on data integration and so not all the issues have been identified. A second factor is that there seem to be two distinct steps in combining data. The first involves the integration of data from different job portals, while the second involves integration of the resulting composite micro data source with the JVS to produce a final statistical output. These seem to be two quite distinct steps and it is not clear where this should fit within this framework. Reid et al. (2017), propose a third phase 3 that focuses on potential errors resulting from the creation of final estimates from the composite micro-data. This may be where integration with survey data should be considered.

4.2.2 UNECE Big Data Quality Framework

This framework has three "hyper-dimensions"; the source, the metadata, and the data, with quality dimensions nested within each of the hyper-dimensions. These dimensions apply to the three phases of the business process: input, throughput, and output. An assessment was made against in the input phase but time constraints prevented a full assessment on-line job advertisement data against the whole framework.

[9] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Virtual_Sprint_1_February_2017

The UNECE framework addresses more qualitative type issue such as data usage (of the data owner), readability, reliability, and availability. The framework thus provides a means of capturing issues such as, whether certain types of data (e.g. pdf files and images) scraped from job portals can be used, the primary purpose of the data being available (i.e. advertising jobs) and risks around continued availability (e.g. a job portal going out of business).

4.2.3 Conclusion:

In general, the UNECE framework seems to be more intuitive and it is and so could be a more useful starting point for identifying issues on quality on on-line advertisement data. However, the Statistics New Zealand Quality Framework is designed to support a total survey error approach which could further deepen the accuracy and selectivity dimension of the UNECE framework. These elements would become more important when considering how on-line job advertisements could be moved into statistical production.

4.2 Measuring Coverage:

As previously discussed, one of the fundamental quality issues in using on-line job advertisement data is that not all job vacancies are advertised on-line. Understanding these issues of coverage has been a key focus for this pilot and three distinct approaches have been indentified for trying to better understand and measure these differences:

i. Measuring use of advertising channels via the JVS: In 2015, Germany included a supplementary module in their JVS on advertising channels used by employers. Slovenia done something similar very recently and the results are presented in the report annex. The results from these surveys can be used to estimate the proportion of employers using on-line advertising channels as well as the proportion and type of jobs.

ii. Matching JVS reporting units to company counts from on-line sources: This has been the main focus of the UK pilot. This approach enables JVS vacancy counts by company to be compared directly with those from on-line sources, thus providing similar measures of coverage. One of the other benefits of this approach is that the different results by job portal can be used to identify which large job portals (or other aggregate sources) have the best coverage and therefore, were efforts on accessing data should be focused.

iii. Comparison of aggregates: This approach involves comparing JVS aggregates by industry sector with the equivalent taxonomies from job portals. This approach has been used by Germany (DESTATIS), who do not have access to JVS micro data. While this is quite a straightforward approach main problem here is that each portal has different taxonomies, which are only approximately comparable with the survey.

4.3 Matching:

Four of the country pilots have explored the matching of on-line job ads with their own JVS micro data as a means of starting to understand coverage issues. The results have been somewhat mixed. Matching rates between data from the Swedish Employment Agency (SEA) and the JVS has been very high since SEA job ads contain the same organisation identifier as used in the JVS. In contrast, other matching has been done on company name only, which has proved much more difficult. Common problems include use of abbreviated names, trading names rather than the legal enterprise name, and misalignment between the company names and the JVS reporting unit. Greece also has some specific issues with the mixed use of Greek and Latin scripts in different data sets. The UK have tried to circumvent some of these issues by matching company name from two data sets, the web scraping framework is designed to target the specific reporting units from the job vacancy survey. This is a manually intensive approach but produces good quality results.

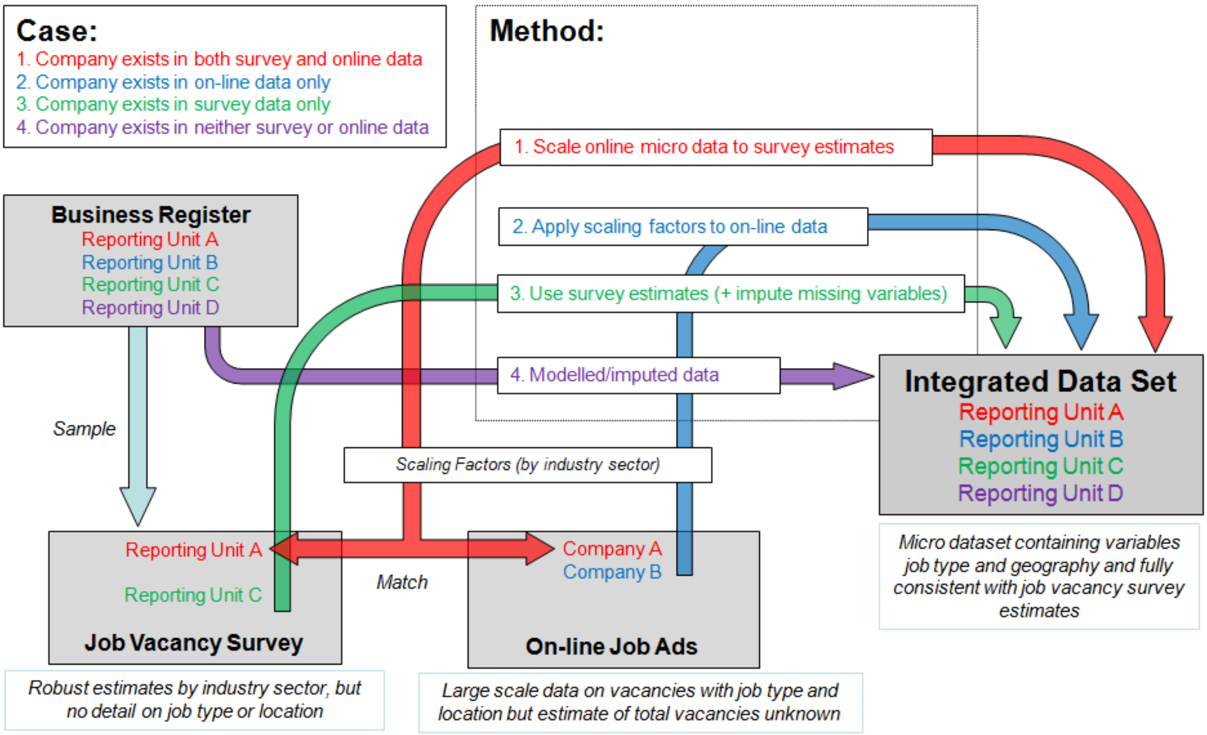
A major issue in trying match JVS reporting units to company names in job portals is that many jobs are advertised through private employment agencies and the employer is not usually identified in the advertisement. In some cases, there may be clues in the job ad about the type of business, or its location. Also, if matching counts between the JVS and direct employer counts from the on-line sources, then any shortfalls in the on-line data may provide further clues as to which employers and what type of jobs are being advertised through employment agencies. Work undertaken by Slovenia reveals that slightly less than 10 per cent of job offers are made through employment agencies. However, this proportion may well be higher in other countries - In 2015, the UK had 14,280 enterprise units classified as activities of employment placement agencies[10]. This will be an important area of further exploration.

4.4 Outline approach to data integration:

One potential long-term approach to statistical production could be to develop an integrated micro data set combining both online data sources and the JVS so that the outputs were fully consistent with aggregates from the JVS. An outline of how this might work is presented in Figure 1. The aim is to have an integrated dataset of job vacancies that is consistent with the survey data but contains the additional granularity of jobs that are advertised online. This would enable statistics about the job type (e.g. occupation) and geographic information about job vacancies to be published. The idea would be to link the survey and the on-line sources by enterprise/reporting unit and then calibrate the on-line micro data to the survey aggregates.

[10] https://www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/datasets/businessdemographyreferencetable

Figure 1 Outline Approach to Data Integration


Case:
1. Company exists in both survey and online data
2. Company exists in on-line data only
3. Company exists in survey data only
4. Company exists in neither survey or online data
Method:
1. Scale online micro data to survey estimates
2. Apply scaling factors to on-line data
3. Use survey estimates (+ impute missing variables)
4. Modelled/imputed data
Business Register
Reporting Unit A
Reporting Unit B
Reporting Unit C
Reporting Unit D
Sample
Scaling Factors (by industry sector)
Integrated Data Set
Reporting Unit A
Reporting Unit B
Reporting Unit C
Reporting Unit D
Micro dataset containing variables job type and geography and fully consistent with job vacancy survey estimates
Reporting Unit A
Reporting Unit C
Job Vacancy Survey
Match
Company A
Company B
On-line Job Ads
Robust estimates by industry sector, but no detail on job type or location
Large scale data on vacancies with job type and location but estimate of total vacancies unknown

There are then different possibilities in terms of the different entities in different data sets and how these could be combined:

Case 1: A match exists between reporting unit and big data source. Data are calibrated to the big data source using the survey estimates.

Case 2: The reporting unit exists in the big data source but not the survey. The scaling factors are calculated and applied for each industry group based on the outputs of Case 1.

Case 3: The reporting unit exists in the survey but not the big data source. These will most likely be small business with few or no vacancies and would simply use the survey estimates. There may be a need to impute some values. Location would be easy as nearly all would be single location enterprises. Job type / occupation data could be imputed based on occupations data derived from corresponding industry codes from Cases 1 and 2.

Case 4: Neither on-line or the survey. As with Case 3 most will be small businesses with few if any vacancies. Data could be modelled based on data derived from the other 3 Cases.

5. Statistical Outputs

This pilot is still some way from being able to produce even experimental statistical outputs. However, we are confident that there will be concrete experimental outputs by the end of SGA-2.

6. Future Perspectives

While some progress has been made in this pilot, there is a way to go, even just to produce experimental estimates using on-line job data. There are many areas that could be explored further including:

- Improved methods for matching with survey reporting units and for developing measures of coverage.
- Improved deduplication methods.
- Methods for identifying other issues in aligning target concepts and measures (e.g. ghost vacancies, multiple vacancies per advertisement).
- Further exploration into web scraping enterprise websites.
- Obtaining data from employment agencies or otherwise adjusting for job advertisements advertised by employer.
- Enrichment of web scraped data including by geography and occupation.
- Use of NLP techniques to capture other information from the full job description (e.g. skills, educational requirements).
- Alignment of definitions.
- Integration of multiple on-line sources.
- Further work on integrating survey and on-line data and estimatation methods.

In short, there is a huge amount still to do and it is not feasible to tackle all these aspects by the end of the ESSNet. Therefore there is a need to prioritise work that will deliver the most benefit.

It is clear that the effort involved in building robots to scrape job portals, maintaining them and developing systems to process these data for statistical purposes, particularly for countries with large labour markets, is very considerable. Furthermore, interest in on-line job advertisement data extends beyond the official statistics community - there are other public organisations that could also make use of these data. It makes more sense therefore, that efforts to obtain these data are coordinated and performed at scale so the duplication of effort is minimised.

The project underway by CEDEFOP to develop a framework for capturing on-line job vacancies for all EU member states is a promising development, and there is an opportunity to collaborate and influence the direction of this work. NSIs have particular skills around understanding and assessing quality and they also have access to JVS data that can be used to benchmark the quality of this kind of data.

For this reason, it is proposed that this work package does not simply aim to replicate the kinds of activities that other organisations are already doing, in particular, around the transformation of raw web scraped data about on-line job advertisements into data suitable for analysis. Instead, efforts should be focused on partnering with data owners and identifying areas were NSIs can add the value to these existing data, particularly around quality, use of standard taxonomies, exploitation of survey data, business registers and other data.

References:

Carnevale A., Jayasundera T., Repnikov D., 2014, “Understanding on-line job ads data: A Technical Report”, Georgetown University; Available at: https://cew.georgetown.edu/wpcontent/uploads/2014/11/OCLM.Tech_.Web_.pdf (Accessed 25 June 2017:

Stateva, G., ten Bosch, O., Maslankowski, J., Righi, A., Scannapieco, M., Greenaway, M., Swier, N., Jansson, I., Wu, D., 2016, “Legal Aspect to Web Scraping of Enterprise Websites”, Eurostat; Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/a0/WP2_Deliverable_2_1_15_02_2017.pdf (Accessed 29 June 2017)

Ketner, 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; Available at: http://ec.europa.eu/eurostat/documents/3888793/5847769/KS-RA-10-027-EN.PDF/87d9c80c-f774-4659-87b4-ca76fcd5884d (Accessed 24 Oct 2016)

Körner, T., Rengers, M., Swier, N., Metcalfe, E., Jannson, I., Wu, D., Nikic, B., Pierrakou, C., 2016, “Inventory and qualitative assessment of job portals, Eurostat; Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_draft_v5.docx (Accessed: 1 November 2016)

Swier, N., Metcalfe E., Jansson, I., Wu, D., Nikic, B., Pierrakou, C., Körner, T., Rengers, M., 2016, “Interim Technical Report (SGA-1)”, Eurostat; Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/64/WP1_Deliverable_1_2_final.pdf (Accessed: 31 July 2017)

Annex A : Germany

1. Assessment of Job Portals in Germany

As a result of a first general analysis of the job market by using web search engines undertaken in January 2016 there were found some useful websites which have rankings or assessment analysis of job portals. In Germany, there are several worth mentioning:

(i) deutschlandsbestejobportale.de

This web site contains ranking lists of job portals or job search machines from 2010-2015. The initiators of the test called “DeutschlandsBesteJobportale" (best job portals of Germany) are ICR, Institute für Competitive Recruiting (competitiverecruiting.de) and the joint project CrossPro Research (crosspro-research.com). The latter project is a corporation project of Cross Water Systems and PROFILO Rating GmbH.

(ii) crosswater-job-guide.com

The web sites crosswater-job-guide.com and crosswater-systems.com belong to the company Crosswater Systems. According to the company's own information the aim is to assist job seekers by providing web guides on a selected range of topics, including annotated links and web resources, to offer the job seeker a pre-selection of jobs that may be of interest. They provide, among others things an own assessment of online job portals on the sub web site: jobbörsen-kompass.de.

(iii) online-recruiting.net

This web site contains many freely available research results, e.g. description of job portals and job portal rankings. It also names job portal URLs of the following 27 countries: Australia, Austria, Belgium, Bosnia Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, France, Germany, Great Britain, Greece, Hungary, India, Italy, Japan, Jordan, Netherlands, Norway, Poland, Portugal, Romania, Russia, South Africa, Spain and Switzerland.

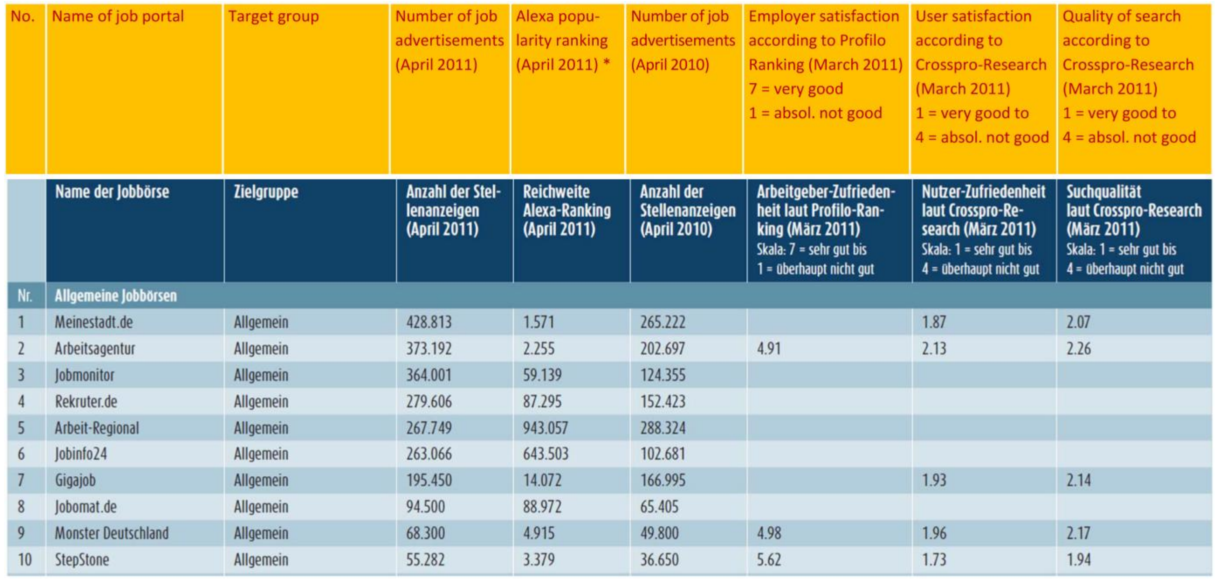
(iv) jobboersen-im-test.de

This website is run by a private individual interested in the topics of online job-placements and -recruiting.

According to the information from *deutschlandsbestejobportale.de* in 2015 there were more than 1600 job portals for the German labour market. *Crosswater-job-guide.com* lists no less than 1794 job portals of various kinds, of which 1088 are earmarked as being currently active. *Jobboersen-im-test.de/* (“Jobbörsen A-Z“) lists altogether 781 job portals (only classified by “target group, profession or sector“, but not by “kind of job portal“). Of these 781 job portals 34 are ranked among the best job portals. *Online-recruiting.net* lists 99 job portals; some of them are not included in the above mentioned 781-list. Similarly, also the 1088-list of *Crosswater-job-guide.com* does not include all the job portals shown in the other lists.

At the end of this task, neither the precise number of job portals nor their importance for the job market was really clear. Figure 2 shows that, depending on the ranking criteria and/or the reference date quite diverging rankings can result. Besides the number of job advertisements, there are four other ranking criteria, namely the Alexa popularity ranking, the employer satisfaction according to Profilo ranking, the user satisfaction according to Crosspro-Research and the quality of search according to Crosspro-Research.

Figure 2: Compilation of ranking lists



No.	Name of job portal	Target group	Number of job advertisements (April 2011)	Alexa popularity ranking (April 2011) *	Number of job advertisements (April 2010)	Employer satisfaction according to Profilo Ranking (March 2011) 7 = very good 1 = absol. not good	User satisfaction according to Crosspro-Research (March 2011) 1 = very good to 4 = absol. not good	Quality of search according to Crosspro-Research (March 2011) 1 = very good to 4 = absol. not good
	Name der Jobbörse	Zielgruppe	Anzahl der Stellenanzeigen (April 2011)	Reichweite Alexa-Ranking (April 2011)	Anzahl der Stellenanzeigen (April 2010)	Arbeitgeber-Zufriedenheit laut Profilo-Ranking (März 2011) Skala: 7 = sehr gut bis 1 = überhaupt nicht gut	Nutzer-Zufriedenheit laut Crosspro-Research (März 2011) Skala: 1 = sehr gut bis 4 = überhaupt nicht gut	Suchqualität laut Crosspro-Research (März 2011) Skala: 1 = sehr gut bis 4 = überhaupt nicht gut
Nr.	Allgemeine Jobbörsen							
1	Meinestadt.de	Allgemein	428.813	1.571	265.222		1.87	2.07
2	Arbeitsagentur	Allgemein	373.192	2.255	202.697	4.91	2.13	2.26
3	Jobmonitor	Allgemein	364.001	59.139	124.355			
4	Rekruter.de	Allgemein	279.606	87.295	152.423			
5	Arbeit-Regional	Allgemein	267.749	943.057	288.324			
6	Jobinfo24	Allgemein	263.066	643.503	102.681			
7	Gigajob	Allgemein	195.450	14.072	166.995		1.93	2.14
8	Jobomat.de	Allgemein	94.500	88.972	65.405			
9	Monster Deutschland	Allgemein	68.300	4.915	49.800	4.98	1.96	2.17
10	StepStone	Allgemein	55.282	3.379	36.650	5.62	1.73	1.94

*This ranking is according to the Internet statistics provider Alexa (www.alexa.com), a subsidiary of Amazon.

Source: personalmagazin 06 / 11

This ambiguity leads to the requirement of a second action with the URLs identified in the first action. On the basis on the number of job advertisements as indicated in the results of the first action, **56** job portals were selected and examined in more detail. Table 3 shows these portals alphabetically sorted and classified by one of three types of job portal, (1) job search engines, (2) general job portals or (3) specialized online job portals. Categories two and three include both job boards and hybrid portals.

More details for the 25 job search engines are shown in Table 4; as well as the website name and the URL the owner of the portal, the company name and their registered office address are shown. These are ranked by the given number of job advertisements in Germany, where available. This was also done for the 16 specialized job portals (Table 3). An additional column is included here, describing the area of specialization.

Table 3: Germany - job portals in type (alphabetical order)

No.	Name	URL	Type of job portal
1	Absolventa Jobnet	https://www.absolventa.de/	specific job portal
2	Adzuna	https://www.adzuna.de/	Job search engine
3	arbeiten.de	http://www.arbeiten.de	Job search engine
4	backinjob	http://www.backinjob.de/	Job search engine
5	Betriebs-Berater-Jobs	http://www.betriebs-berater-jobs.de/	specific job portal
6	Careerjet	http://www.careerjet.de/	Job search engine
7	Cesar	http://www.cesar.de/	Job search engine
8	Connecticum	http://www.connecticum.de/	specific job portal
9	Experteer	http://www.experteer.de/	specific job portal
10	fazjob.net	http://fazjob.net/	specific job portal
11	Gigajob	http://de.gigajob.com/index.html	general job portal
12	goodmonday	http://www.goodmonday.de/	Job search engine
13	hotelcareer	http://www.hotelcareer.de/	specific job portal
14	Indeed	http://de.indeed.com/	Job search engine
15	Jobanzeigen.de	https://www.jobanzeigen.de/	Job search engine
16	Jobbörse Bundesagentur für Arbeit	http://jobboerse.arbeitsagentur.de	general job portal
17	Jobbörse.com	https://www.jobbörse.com/	Job search engine
18	Jobbörse.de	https://www.jobbörse.de/	Job search engine
19	Jobcluster	https://www.jobcluster.de/	general job portal
20	Jobkralle	http://www.jobkralle.de/	Job search engine
21	Jobkurier	http://www.jobkurier.de/	specific job portal
22	Jobleads	https://www.jobleads.de/	specific job portal
23	Jobmonitor	http://de.jobmonitor.com/	general job portal
24	Jobrapido	http://de.jobrapido.com/	Job search engine
25	JobRobot	http://www.jobrobot.de/	Job search engine
26	JobScout24 / Jobs.de	http://www.jobs.de / www.jobscout24.	general job portal
27	JobStairs	https://www.jobstairs.de/	general job portal
28	Jobsterne	http://www.jobsterne.de/	specific job portal
29	Jobsuma	http://www.jobsuma.de/	Job search engine
30	Jobturbo	http://jobturbo.de/	Job search engine
31	Jobvector	http://www.jobvector.de/	specific job portal
32	Jobware	http://www.jobware.de/	specific job portal
33	Jobworld	http://www.jobworld.de/	Job search engine
34	Jooble	http://de.jooble.org/	Job search engine
35	Kalaydo	http://www.kalaydo.de/jobboerse/	general job portal
36	Kimeta	http://www.kimeta.de/	Job search engine
37	LinkedIn	https://de.linkedin.com/job/	general job portal
38	Meine Stadt.de	http://jobs.meinestadt.de/deutschland	general job portal
39	Monster	http://www.monster.de/	general job portal
40	njobs	https://www.njobs.de/	Job search engine
41	Online-Jobs.de	http://www.online-jobs.de/	specific job portal
42	Opportuno	http://www.opportuno.de/	Job search engine
43	Placement24	http://www.placement24.com/de/	specific job portal
44	Rekruter	http://www.rekruter.de/	general job portal
45	Renego	https://www.renego.de/	Job search engine
46	Staufenbiel	https://www.staufenbiel.de/startseite.	specific job portal
47	Stellenanzeigen.de	http://www.stellenanzeigen.de/	general job portal
48	stellenanzeigen.net	http://www.stellenanzeigen.net/	Job search engine
49	StepStone	https://www.stepstone.de/	general job portal
50	Süddeutsche Zeitung	http://stellenmarkt.sueddeutsche.de/	general job portal
51	Trovit Jobs	http://de.trovit.com/jobs/	Job search engine
52	Unicum Karrierezentrum	http://karriere.unicum.de/	specific job portal
53	XING	https://www.xing.com/jobs/	general job portal
54	xljob.de	http://www.xljob.de/	Job search engine
55	Yourfirm.de	http://www.yourfirm.de/	specific job portal
56	yovadis	http://www.yovadis.de/	Job search engine

Table 4: Germany - job search engines, ranked by number of job advertisements

		Job search engines			
			Owner of the portal		
No.	Name	URL	Company	registered office	job advertisements in Germany
1	Jobbörse.com	https://www.jobbörse.com/	XING AG	Hamburg (DE)	2,500,000
2	Kimeta	http://www.kimeta.de/	kimeta GmbH	Darmstadt (DE)	2,126,977
3	Cesar	http://www.cesar.de/	cesar Internetdienste GmbH	Hamburg (DE)	1,800,000
4	Jobrapido	http://de.jobrapido.com/	Jobrapido Srl	Milan (IT)	1,714,177
5	Jobbörse.de	https://www.jobbörse.de/	CareerNetwork JOBBÖRSE.de GmbH & Co KG	Wiesbaden (DE)	1,677,214
6	Careerjet	http://www.careerjet.de/	Careerjet Ltd	London (UK)	1,325,511
7	Jobkralle	http://www.jobkralle.de/	Webintegration IT Service GmbH	Wien (AT)	1,044,069
8	Renego	https://www.renego.de/	Renego - Nikolay Nikolov	Köln (DE)	1,000,000
9	JobRobot	http://www.jobrobot.de/	JobRobot e.K.	Hamburg (DE)	918,754
10	Trovit Jobs	http://de.trovit.com/jobs/	Trovit Search, S.L.	Barcelona (ES)	761,651
11	Adzuna	https://www.adzuna.de/	Adhunter ltd.	London (UK)	492,951
12	Jooble	http://de.jooble.org/	Ladoburn Europe LTD	Limassol (CY)	418,810
13	Jobworld	http://www.jobworld.de/	Internext GmbH	Karlsruhe (DE)	400,000
14	Indeed	http://de.indeed.com/	Indeed Ireland Operations Limited	US, Amsterdam (NL), Dublin (IE), Düsseldorf (DE), Hyderabad (IN), London (UK), Paris (FR), Sydney (AU), Tokio (JP), Toronto (CA)	389,211
15	Jobanzeigen.de	https://www.jobanzeigen.de/	classmarkets gmbH	Berlin (DE)	334,539
16	Opportuno	http://www.opportuno.de/	Opportuno GmbH	Erlangen (DE)	312,377
17	Jobsuma	http://www.jobsuma.de/	JOBSUMA GmbH	Köln (DE)	286,119
18	Jobturbo	http://jobturbo.de/	karriere.de / lookas GmbH	Köln (DE)	267,897
19	stellenanzeigen.net	http://www.stellenanzeigen.net/	JobValue GmbH	München (DE)	204,637
20	yovadis	http://www.yovadis.de/	Diekmeyer Medienagentur	Rödental (DE)	123,753
21	xljob.de	http://www.xljob.de/	HR4YOU Solutions GmbH & Co. KG	Großefehn/Timmel (DE); Ebermar	30,249
22	arbeiten.de	http://www.arbeiten.de	StepStone Deutschland GmbH	Düsseldorf (DE)	???
23	backinjob	http://www.backinjob.de/	Ideenkraftwerk GmbH	Herford (DE)	???
24	njobs	https://www.njobs.de/	njobs Ltd	London (UK)	???
25	goodmonday	http://www.goodmonday.de/	platron	Berlin (DE)	???

Table 5: Germany - specific job portals, ranked by number of job advertisements

		Specific job portals				
			Owner of the portal		Number of job advertisements	
No.	Name	URL	Company	registered office		Area
1	Experteer	http://www.experteer.de/	Experteer GmbH	München (DE)	175,046	specialists and managers
2	Jobleads	https://www.jobleads.de/	JobLeads GmbH	Hamburg (DE)	83,090	specialists and managers
3	Placement24	http://www.placement24.com/de/	Placement24 GmbH	Düsseldorf (DE)	75,395	specialists and managers
4	Jobsterne	http://www.jobsterne.de/	Matthaes Verlag GmbH	Stuttgart (DE)	46,392	hotel and catering industry
5	Connecticum	http://www.connecticum.de/	connecticum GmbH	Berlin (DE)	34,207	students, alumni, young professionals (SAY´s)
6	Betriebs-Berater-Jobs	http://www.betriebs-berater-jobs.de/	Deutscher Fachverlag GmbH	Frankfurt a. M. (DE)	25,746	legal, tax and financial consulting
7	Jobware	http://www.jobware.de/	Jobware Online-Service GmbH	Paderborn (DE)	15,105	specialists and managers
8	hotelcareer	http://www.hotelcareer.de/	YOURCAREERGROUP GmbH	Düsseldorf (DE)	14,082	hotel and catering industry
9	Absolventa Jobnet	https://www.absolventa.de/	Absolventa GmbH	Berlin (DE)	7,330	students, alumni, young professionals (SAY´s)
10	fazjob.net	http://fazjob.net/	Frankfurter Allgemeine Zeitung GmbH	Frankfurt a. M. (DE)	6,983	specialists and managers
11	Staufenbiel	https://www.staufenbiel.de/startseite.html	Staufenbiel Institut GmbH	Köln (DE); Frankfurt a. M. (DE)	6,117	students, alumni, young professionals (SAY´s)
12	Jobkurier	http://www.jobkurier.de/	CHECKPOINT HRnetworks GmbH	Bretten (DE)	4,891	specialists and managers
13	Online-Jobs.de	http://www.online-jobs.de/	CareerNetwork JOB-BÖRSE.de GmbH & Co KG	Wiesbaden (DE)	2,800	information technology
14	Jobvector	http://www.jobvector.de/	Capsid GmbH	Düsseldorf (DE)	1,122	natural scientists, medical doctors and engineers
15	Yourfirm.de	http://www.yourfirm.de/	yourfirm GmbH	München (DE)	???	mid-sized sector
16	Unicum Karrierezentrum	http://karriere.unicum.de/	UNICUM GmbH & Co KG	Bochum (DE)	???	students, alumni, young professionals (SAY´s)

Finally, the 15 general job portals listed in Table 3 were assessed in more detail.

Table **6** shows these portals ranked by number of job advertisements - column B contains the numbers used in Table 3. The name and URL were supplemented by information on the owner and the number of job advertisements in Germany (see column D and G in

Table **6**). Column E shows whether the general job portals are job boards or hybrid portals (although in some cases this distinction could not be made. Column F shows whether the portal cooperates with partners, based on information that was publicly available. According to the purpose of the partnerships two different categories were identified: range partners for own job advertisements and hit list partners for job search results. The aim in the first case is to extend the range and the scope of its own job advertisements by allowing partners to publish them on their web sites. In the second case the aim is to expand the portal's own hit list for the job seeker by reporting additional job advertisements from other portals.

To achieve a comparable assessment of the size of the general job portals, an attempt was made to find out the number of job advertisements that are no older than 30 days (column H in

Table **6**). This analysis has been conducted by the use of web scraping techniques. Unfortunately, due to technical reasons, it was not always possible to get this information. For the assessment of hybrid job portals an additional web scraping result was needed: the number of job advertisements *not* older than 30 days *and not* carried over from other portals or enterprise websites. The results are shown in column I of

Table **6**. A comparison between column G and column I shows that the ranking order changes when these results are taken into account.

The comprehensiveness of the structured information of the job descriptions was another element of interest for the assessment of general job portals. That is the reason why it was examined how many descriptive job criteria are available as structured information (column J and K). The more there is, the better, as unstructured information requires the development of complex text mining technologies which is a highly complex task that requires considerable effort. As shown in table 4, the general hit lists of the job portals rarely provide more than four descriptive job criteria (usually including the title of the job, the publication date of the job advertisement, the location and the name of the employer). Some portals provide up to 9 criteria on a second level (accessed when clicking on the job advertisement in the general hit list), while others just provide the job advertisement in unstructured form without additional criteria (indicated by a "0" in column K of table 4).

The selection of the job portals for the further course of the pilot study under WP 1 was mainly based on the size of the job portal (i.e. the number of job advertisements). At the same time, both job boards and hybrid portals were included[11]. As one of the hybrid portals, the job portal of the German Federal Employment Agency (Jobbörse Bundesagentur für Arbeit) needs to be included as it combines the by far largest job board with job advertisements from around 100 partners and job advertisements detected by a job robot on more than 400,000 enterprise web sites. This approach is at the same time of large interest regarding the methodology developed in the pilot study. Apart from that another large, but privately run, hybrid portal (gigajob) was selected, together with the two largest job boards (Stepstone and Monster).

[11] 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.

As a result of the additional assessment analyses shown in

Table **6** for the ESSnet pilot study the following portals were selected:

1. Jobbörse Bundesagentur für Arbeit (hybrid)
2. Gigajob (hybrid)
3. StepStone (job board)
4. Monster (job board)

Table 6: Germany – 15 general job portals ranked by number of job advertisements and further differentiated between job boards and hybrid portals

					General job portal					
Numbers		Name of job portal	Owner of the portal	Type of job portal	Partner	Number of job advertisements in Germany			Number of descriptive job criteria from structured information	
according to column G	referring to table 1			- job board - hybrid	- range partners for own job advertisement - hit list partners for job search results	altogether	not older than 30 days	own job ad AND not older than 30 days	available on general hit list	available by clicking on the ad link
A	B	C	D	E	F	G	H	I	J	K
1	16	Jobbörse Bundesagentur für Arbeit (public employment agency) http://jobboerse.arbeitsagentur.de	Bundesagentur für Arbeit (Federal Employment Agency)	hybrid	around 100 range partners and hitlist partners sharing (parts of) the job advertisements (upon the request of the employer) in addition a job robot captures job advertisements on more than 400,000 enterprise websites	1,083,929	NA	NA	5	9
2	53	XING https://www.xing.com/jobs/	XING AG	hybrid	interface to the job advertisement database of the Federal Employment Agency (hit list partner)	570,797	NA	NA	4	6
3	11	Gigajob http://de.gigajob.com/index.html	Netzmarkt Internetservice GmbH & Co. KG	hybrid	16 hit list partners (most of them job search engines)	531,112	417,229	212,157*	4	4
4	37	LinkedIn https://de.linkedin.com/job/	LinkedIn Ireland	hybrid	interface to the job advertisement database of the Federal Employment Agency (hit list partner)	488,098	486,791	NA	4	8
5	38	Meine Stadt.de http://jobs.meinestadt.de/deutschland/stellen	meinestadt.de GmbH	hybrid	job advertisements from established business partners and directly from enterprise websites => hit list partners	466,680	360,086	6,287	4	0
6	44	Rekruter http://www.rekruter.de/	FM-Studios GbR www.fm-studios.de	hybrid	interface to the job advertisement database of the Federal Employment Agency (hit list partner)	NA	NA	NA	4	5
7	26	Jobs.de / JobScout24 http://www.jobs.de www.jobscout24.de	CareerBuilder Germany GmbH	hybrid	a lot of advertisements from personnel recruiting persons and temporary work agenies; 9 range partners	115,867	115,194	28,618	4	0
8	23	http://de.jobmonitor.com	Harald Stückler	hybrid	also hit list results form "Monster" and "youfirm"	105,905	105,905	50,130	4	0-4
9	49	StepStone https://www.stepstone.de	StepStone Deutschland GmbH	job board		61,119	60,938	60,938	4	6
10	19	Jobcluster https://www.jobcluster.de	Jobcluster Deutschland GmbH	hybrid	cooperation partner of the Federal Employment Agency (hit list partner)	43,741	NA	NA	6	8
11	39	Monster http://www.monster.de	Monster Worldwide Deutschland GmbH	job board	merger between Monster and Jobpilot; 167 range partners	39,213	NA	NA	4	0
12	27	JobStairs https://www.jobstairs.de	milch & zucker - Talent Acquisition & Talent Management Company AG	job board	8 "target-group-specific" partnerships 53 enterprises; this job board operates under "Top Company Portal"	25,900	NA	NA	4	0
13	47	Stellenanzeigen.de http://www.stellenanzeigen.de	stellenanzeigen.de GmbH & Co. KG	job board	350 range partners; including well-known meta search engines	9,029	9,029	9,029	4	0
14	50	Süddeutsche Zeitung http://stellenmarkt.sueddeutsche.de	Süddeutscher Verlag	job board		8,972	8,972	8,972	4	0
15	35	Kalaydo http://www.kalaydo.de/jobboerse/	Kalaydo GmbH & Co. KG	job board	range partners: 51 daily newspapers and 7 more advertising papers	6,063	5,958	5,958	4	0

* According to an estimation based on web-scraped data, among these job advertisements designated as „own“, there are more than two-thirds that were original posted on the job board of the public employment agency (no. 1 according to column A).

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) 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 volume of information obtained from a database, the smaller the likelihood of legal concerns to web scrape this information. As a result, it was concluded that legal restrictions were not a major issue for the pilot study, but that for any larger scale production of statistics consent should be obtained from the portal owner, not least to avoid breaks in time series due to lacks in data availability.

Besides the possible legal constraints there may also be technical restrictions. Some portals limit their hit lists for job search results. For example, search results of the job board Monster are limited to 1000 job advertisements. This may be another reason for developing partnerships with job portal owners. In this context answers to the following questions are needed.

1. Can a contract with a job portal owner be made if there is no legal basis for data sharing?
2. Are official statistics allowed to offering anything in return for supplying data?
3. Is the practice fully in line with the European Statistics Code of Practice?
4. How long should the contractual period be?
5. How should these contracts be managed in context of changing business models, market shares and a desire to engage with actual or potential competitors?
6. Can cooperation for statistical purposes be considered not as a burden but as a positive confirmation of the portal owner's own high market relevance provided with a quality label by 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 to obtain a 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, which is characterised by large volumes of data.

3. Job vacancy data available from the Federal Employment Agency and the Institute of Employment Research

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 an online job portal and job robot that scrapes vacant posts from enterprises, which are used for some limited statistical purposes.

Prior to the European regulation 453/2008, Germany had a comprehensive annual job vacancy survey that ran during the 4th quarter of each year. 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., 2015).

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 Moczall et al. 2015; 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.

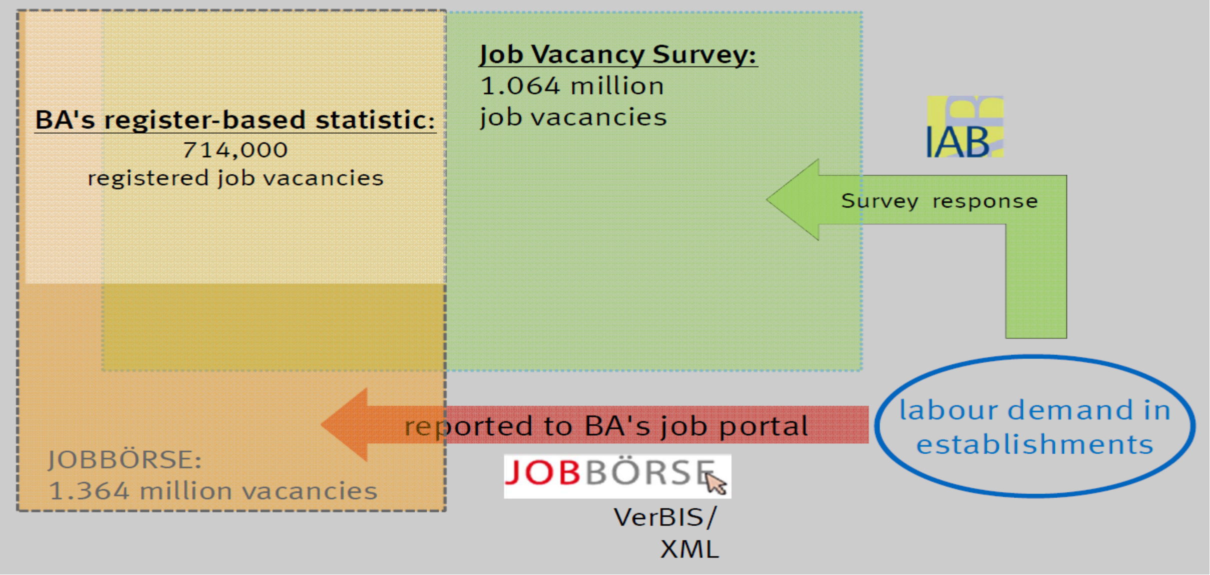
The data of the job vacancy survey which are available on http://www.iab.de/de/befragungen/stellenan
gebot.aspx: in 2017q1 there were found 1.064 million job vacancies. According to the information supplied by the FEA's job portal (https://jobboerse.arbeitsagentur.de) there are 1.364 million jobs, whereas the FEA (see https://statistik.arbeitsagentur.de/) has 714,000 registered job vacancies (both figures as of 19 June 2017).

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).

Figure 3 shows the above mentioned data sources of the Federal Employment Agency. The data of the job vacancy survey which are available on http://www.iab.de/de/befragungen/stellenangebot.aspx: in 2017q1 there were found 1.064 million job vacancies. According to the information supplied by the FEA's job portal (https://jobboerse.arbeitsagentur.de) there are 1.364 million jobs, whereas the FEA (see https://statistik.arbeitsagentur.de/) has 714,000 registered job vacancies (both figures as of 19 June 2017).

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).

Figure 3: Data sources of Federal Employment Agency with latest figures (June 2017 and first quarter 2017 respectively)



4. First results on quality of job portal data

Identifying duplicates

An exploratory study on de-duplication 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 de-duplication procedures. Furthermore, it was concluded that de-duplication 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 de-duplication 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 de-duplication issues.

Comparison job portal data with results of the job vacancy survey (JVS)

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.

The business register in Germany

In the European Union every Member State manages its own national statistical business register. The German business register (abbreviation BR or URS) is a regularly updated database, which contains structured information about enterprises. The main statistical units are enterprises and their local units, legal units of the enterprises and enterprise groups (including both resident and multinational). The business register also holds administrative data and other business characteristics such as taxable turnover and the number of employees subject to social insurance.

The main two sources of the German business register are first monthly data of the Federal Employment Agency and second monthly data of the financial administration (turnover). Additional data sources are data of the Federal Central Tax Office (on tax groups), data of the Chamber of Crafts, information from the trade register (annual data), from commercial data providers for the maintenance of enterprise groups, data from the EuroGroups Register (EGR) and information from different surveys.

In 2015, there was a total of 3.5 million enterprises with 3.7 million local units in Germany (sections B-N and P-S of economic activity). Most enterprises have fewer than ten persons employed and these enterprises accounted for 91% of the total stock of enterprises, whereas enterprises with more than 250 persons employed accounted for almost half of the turnover.

As mentioned above business register data can be helpful for verification of job portal data quality. For this purpose it is necessary to match job portal data with business register data and to get additional information on specific enterprise characteristics (e.g. economic activity).

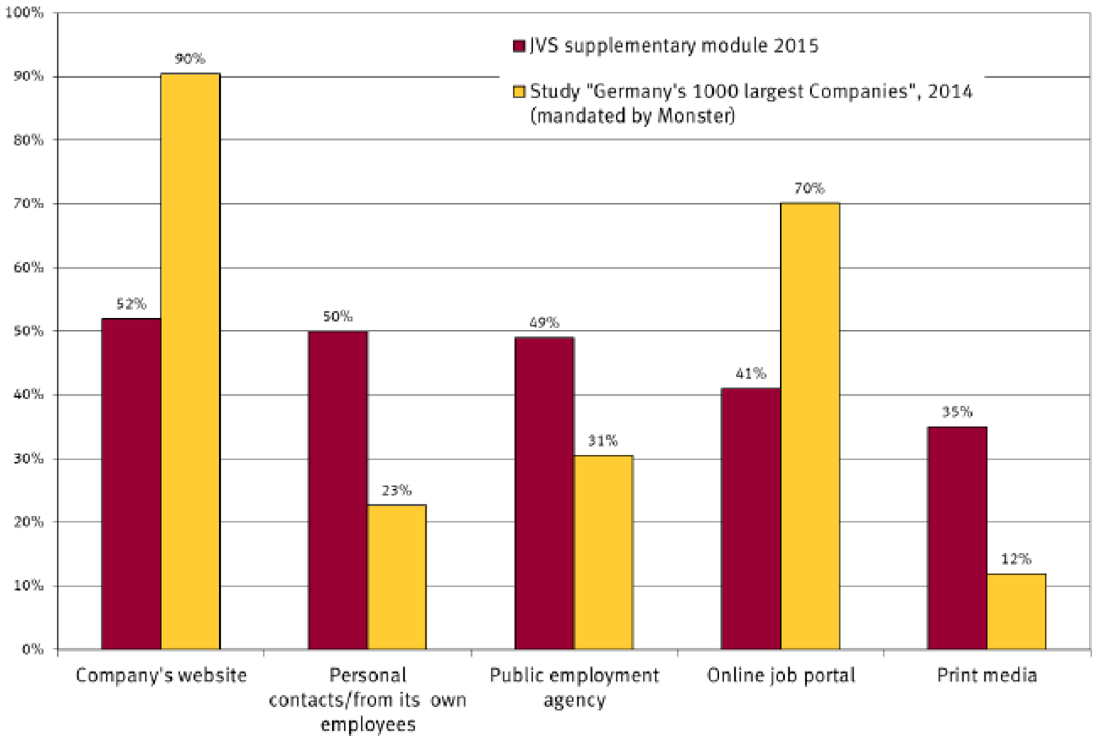
There have been some studies on micro data matching in the past. In previous studies micro data from structural business statistics, business demography, inward FATS and the ICT survey were linked with the business register using the so called URS ID (for further information see Jung / Käuser, 2016). The URS ID is an ID number of every legal unit in the business register. This ID serves as an identifier for linking different structural statistics and also to the business register. The resulting matching rates are quite high at around 99 %. This is due to the fact that the business register serves as a sampling frame for structural business statistics.

The population used for the sampling originates from the currently available address stock of the Federal Employment Agency (FEA) which includes all establishments with at least one employee subject to social insurance contributions. So the FEA's address stock is on the one hand the main data source for the business register and on the other hand the basis of the sampling of the job vacancy survey. As a consequence, statistical units in the job vacancy survey could more or less be seen as a kind of subsample of the business register. This is not the case for job portal data as the URS ID cannot be used. The feasibility of matching using other variables (e.g. company name, address requires further investigation.

5. Recruiting channels used by employers

There are some studies with data on the recruiting channels used by enterprises in Germany. Some of this information was 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.

Figure 4: Recruiting Channels used by German Enterprises



6. Conclusions and next steps

The important characteristics of the German circumstances in the context of job vacancy statistics are as follows:

i) <u>Extremely high number of job portals:</u>
There are at least 1,600 job portals in Germany. Most of them are hybrid portals having hit list partners as well as range partners.

ii) <u>Frequently changes in the market:</u>
There is a frequently change in the market. This concerns not only the business models (e.g. pay per ad, pay per click, matching services) of the job portals but also to a frequently change in the ranking of job portals. It can be assumed that both aspects are mutually dependent.

iii) <u>Restrictions of the information available :</u>
Many hit lists are restricted (e.g. no more than 1000 ads displayed). This case leads to the need for an agreement between statistical office and job portal owner to get access to the data. Another general problem is that more or less all hit lists provide fairly little structured information (4-6 on average). Sometime it is possible to get further information by using the full job ad. As job ads in Germany mostly follow the corporate design of the particular employer there is often no standardised or structured information available.

iv) Importance of Federal Employment Agency:

The great importance of the Federal Employment Agency FEA in the area of job vacancy statistics leads to a special situation in Germany compared to other countries. The FEA is not only responsible for the job vacancy survey but is also the biggest job portal owner in Germany.

Nevertheless Destatis will do further investigation especially in the area of data quality of job portal results. For this purpose the possibilities of matching job portal data with business register data will be investigated during SGA-2. Furthermore the possibilities to get access to the data from the job portal owners will be explored. Apart from investigating the readiness of the portal owners to provide their data for statistical purposes and to questions of data acquisition, 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.

References

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)

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 19 June 2017)

Eurostat (2011): European Statistics Code of Practice, Luxembourg 2011.

European Union (2008). Regulation (EC) No. 453/2008 of the Parliament and of the Council of 23 April 2008 on quarterly statistics on Community job vacancies. Official Journal of the European Union 4.6.2008, L 145/234.

Jung, Sandra / Käuser, Stefanie (2016), CHALLENGE AND POTENTIAL OF LINKING MICRODATA IN BUSINESS STATISTICS, https://www.destatis.de/EN/Publications/WirtschaftStatistik/LinkingMicrodataBusinessStatistics_022016.pdf?__blob=publicationFile; German version published in WISTA 2/2016, p. 95 et seq

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)

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 19 June 2017)

Annex B: Greece

1. Introduction

In Greece, a significant proportion of the jobs are filled via social networking (Villar et al, 2000; Moira et al, 2004). Companies in Greece are using modern technologies and particularly internet sites to communicate with potential job applicants. The online recruitment is attracting a growing number of companies (Galanaki, 2002) as it has low cost and is widely used by job applicants (Anastasiou, S. 2014). Regarding the numbers of job seekers and offers, it is expected to be lower compared to other countries due to the relative small size of the job market (Terzis, V. & Economides, A.A. (2005).

Therefore, the main scope of our research is to investigate the accessibility and data quality of selected job portals, to understand the jobs that are advertised on-line at national level and to explore how these data can be used for estimating the relevant statistics.

The Greek labour market was characterized by a relatively high share of permanent jobs to the total dependent employment on the one hand, and a high share of self-employment that dominated overall employment, on the other hand. However, the depth and duration of the recession have resulted in a severe deterioration of the national labour market and social conditions (ILO, 2014). One in four jobs that existed before the crisis has been lost. The severity of national labour market distress is, also, reflected in the duration of unemployment. More than 70 per cent of the unemployed have been without a job for more than one year.

According to the above mentioned ILO study, the predominance of small-scale entrepreneurial activities is a case in point. Over 96 per cent of enterprises have fewer than 10 employees and employ more than 57 per cent of the total labour force – the highest share in the EU. For many of these firms the opportunities to exploit economies of scale, undertake long-term investments, create decent work opportunities and compete in international markets is limited.

Current survey

The Job Vacancy Survey is conducted on a sample designed on the basis of the Statistical Business Register of ELSTAT. The concepts and definitions of the basic variables used for the compilation of Job Vacancy Statistics are laid down in European Regulations (EC) No 453/2008, (EC) No 1062/2008 and (EC) No 19/2009. (See Annex)

More specifically, for every two-digit code of economic activity a number of enterprises is 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 conducted on a sample of 7,511 enterprises and services.
The data are collected through paper questionnaires.
Non-response is addressed through telephone contacts with the enterprises, reminders sent by fax or e-mail or personal visits to the enterprises.

The statistical population consists of all the enterprises employing at least one (1) employee and belonging to Sections B-S of NACE Rev.2. The reference area is Greece (total) and the reference period of the data on Job Vacancies is one calendar quarter.

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

2. Data Access

Accessing data directly

As a first step of our research, an internet investigation of job portals, job search engines and specialist job sites, was carried out throughout Greece. According to http://www.greek-sites.gr, a site that ranks the Greek sites popularity, 28 job portals with domains ".gr" were found. This list is presented in Annex.

In order to determine on which web sites the pilot study should focus, the major job portals are sorted on the basis of the following criteria: a) the number of advertisements (size); b) monthly visitors (June 2016) and c) the Alexa[12] popularity ranking. The job portals that have been identified as potential of interest for the study after their ranking are shown in Table 1.

Table 1: Greece – List of 14 major job portals

A/A	Name	Number of advertisements (7 June 2016)	Target group	Monthly visitors(X1000) June 2016	Alexa Ranking
1	kariera.gr	1,900	General Job board	855	10,920
2	oaed.gr	_	Recruitment agency	1,400	21,217
3	skywalker.gr	3,300	General Job board	647	27,086
4	proson.gr	1,000	National public website	286	31,709
5	asep.gr	_	National public website	248	73,894
6	jobfind.gr	500	General Job board	163	85,166
7	diorismos.gr	>500	Newspaper Website	172	110,586
8	careernet.gr	350	General Job board	156	112,689
9	randstad.gr	200	Recruitment agency	63	253,594
10	proslipsis.gr	200	Newspaper Website	110	273,287
11	neuvoo.gr	12,000	Job search engine	49	317,627
12	mycarriera.gr	150	Job search engine	22	358,217
13	yourse.gr	3,000	Greek Job search engine	33	615,639
14	jobseeker.gr	50	General Job board	16	873,686

[12] 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.

t is difficult to assess the quality of the different job web sites. However, taking into account the above mentioned analysis the site “Skywalker.gr” was selected for our initial web scraping experiment in the ESSnet pilot study.

The selection of this portal for further analysis was mainly based on the size of the job portal, the Alexa popularity ranking, the comprehensiveness of the structured information of the job vacancy descriptions and the national level coverage. Moreover, it is mentioned as one of the major job vacancy sites operating in Greece (Anastasiou, S. (2014); http://www.greek-sites.gr)

Accessing data from others:

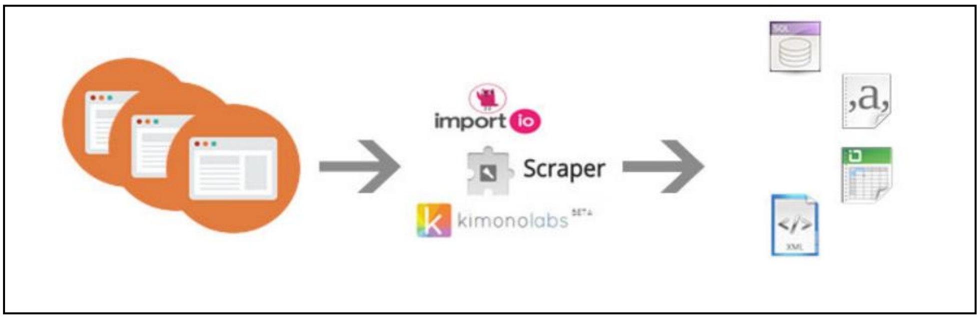
- Government employment agencies

For the public sector, there two main job portals are oaed.gr and asep.gr. These are considered more easily accessible sources. ELSTAT is exploring the possibility to collect data from the public administrative sources (Social Insurance Institute – IKA, “ERGANI” project) in order to enhance the quality of data and reduce the administrative burden of enterprises.

3. Data Handling

ELSTAT has focused on scraping ads directly from job portals. The purpose of our web scraping experiment was to scrape very specific structured information selected from the job portal and for this purpose tools for general scraping purposes(such as import.io and content grabber) were used (See Figure 1).

Figure 5: Web scraping process



The collected data cover the following fields:

- Job category
- Job title
- Job Description (a “snippet” of the job description between 40-60 words)
- Location

- Company name
- Posted date
- Working hours (part time/full time)
- Salary
- Contract type

The selected job portal has links from the job offers to a second level of standardised information, which consists of the full-text of the job advertisements plus further semi-structured information. The first results reveal that there are problems with missing data, especially as regards the field salary and contract type. In addition, there are problems with taxonomy in the Job Category and Location fields.

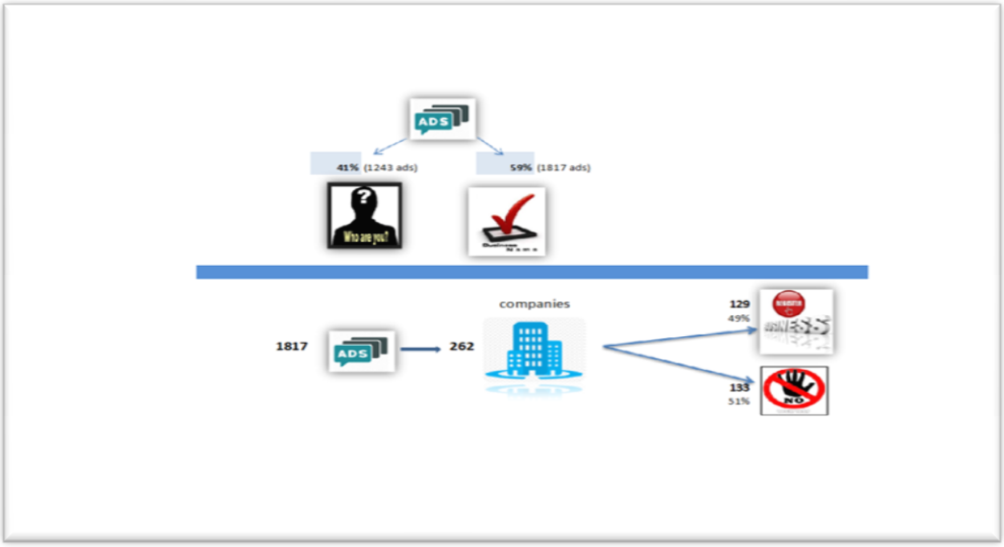
4. Methodology – Statistical Outputs

1st Experiment

In the context of the second sprint, the first experiment undertaken by ELSTAT aimed at exploring to what extent the job portal data covers what is measured by the job vacancy survey. During the two month period from June to August 2016, when the experiment lasted, a “clean” sample dataset of 3060 single advertisements was created from the scraped ads collected.

The first step was to identify the company names from the descriptions in the ads. In 59% of the ads (1817) the company names were identified. For the rest 41% no company name was available. The majority of these particular ads started in a systematic way such as “Leading Company…” or “Well Known Firm...” etc. In total, 262 company names could be matched. The first attempt was to match the companies from the dataset to the ones from the sample of the Job Vacancy Survey (JVS). However, the results were very poor. 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. 49% of these companies were matched (See Figure 2).

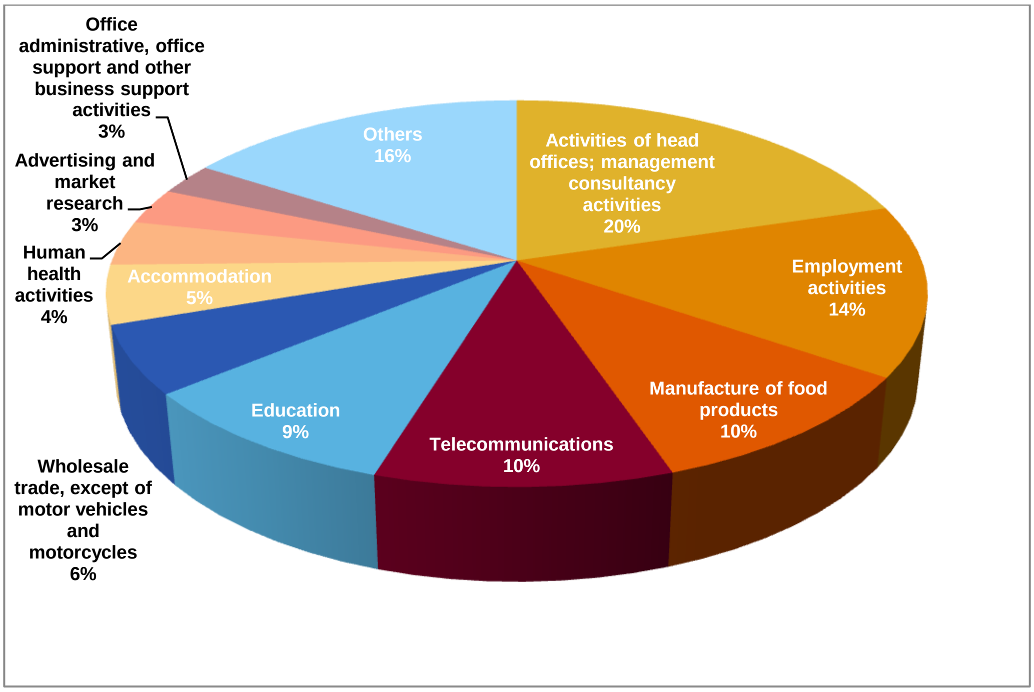
Figure 2: Web Scraping 1st Experiment



The results of the classification of the companies by economic activity reveal that 34% refer to head offices [management consultancy (20%) and employment activities (14%)], 10% manufacture of food products, 10% telecommunications and 9% education (see Figure 3).

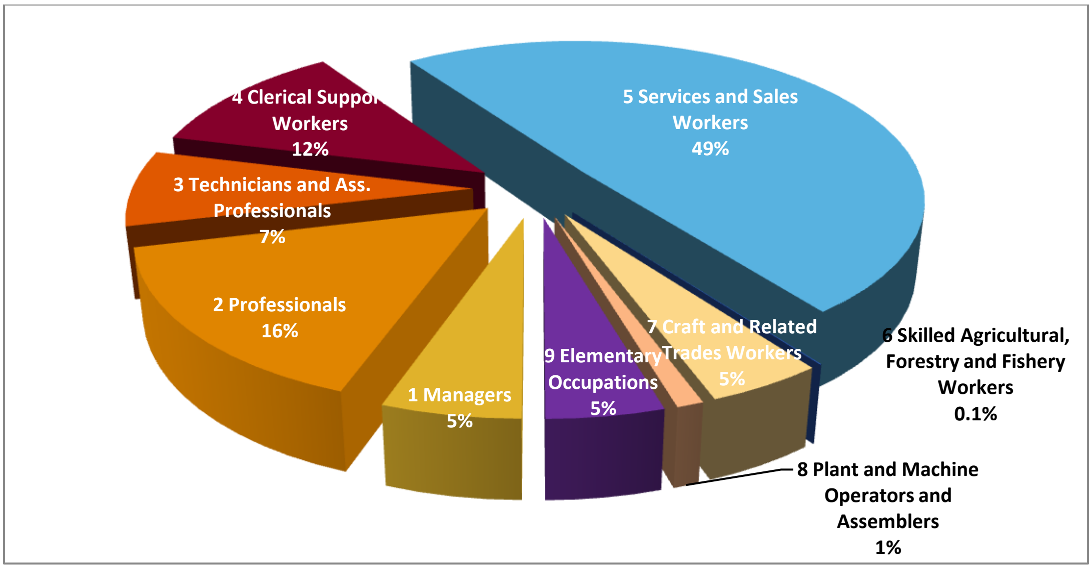
-

Figure 3: Classification of Companies by Economic Activity (NACE rev. 2)



Additionally, job description classification of the ads by major groups (ISCO-08) reveals that almost 50% of the ads are about Services and Sales Workers (see Figure 4). However, in some of the major groups there are low percentages of on line ads.

Figure 4: Job description classification by Major Groups ISCO-08



2^{nd} Experiment

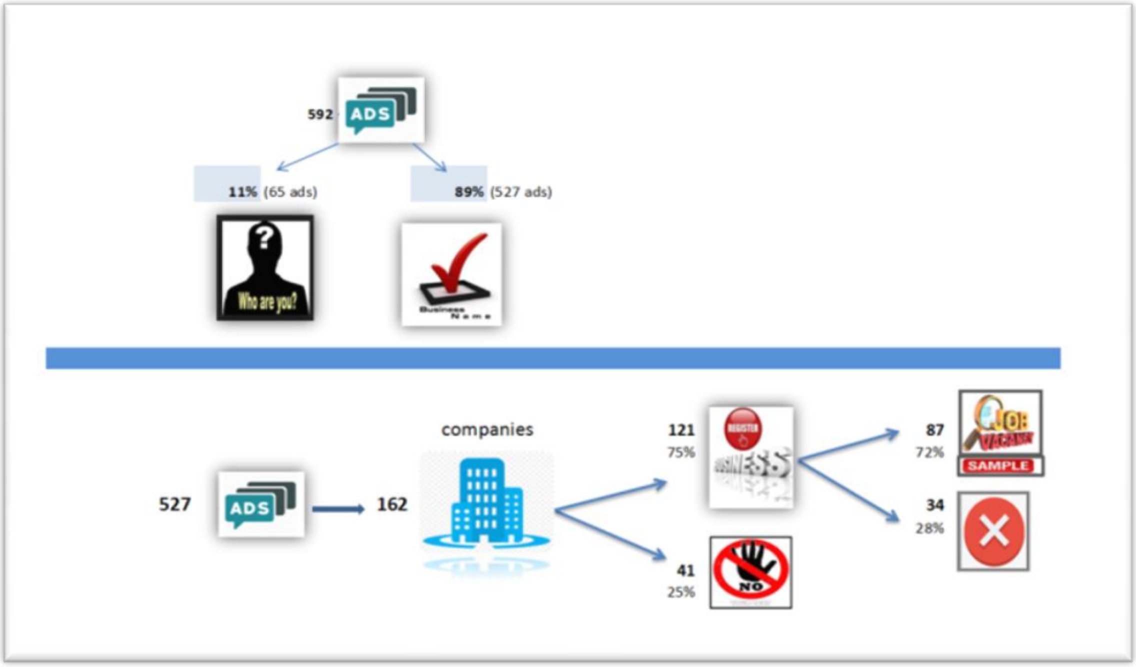
After our first experiment, we decided to focus on IT domain that is most likely for the jobs to be advertised on line.

From November 2016 to January 2017, an API was set up and 925 ads for IT domain were collected. When the dataset was created, the observations that were detected more than once (2%) were eliminated. Moreover, 28% and 7% of job ads referred to non- IT jobs and for working abroad respectively which also were removed, as these were not included in the target population. In this way, a “clean” dataset of 592 single advertisements was created. For the majority of the ads (89%) the company names were identified from the description of the ads, which corresponded to 162 company names. 75% (121) of these were matched with the Statistical Business Register (SBR).

The comparison among the identified companies of the two experiments showed that only 8 companies were common.

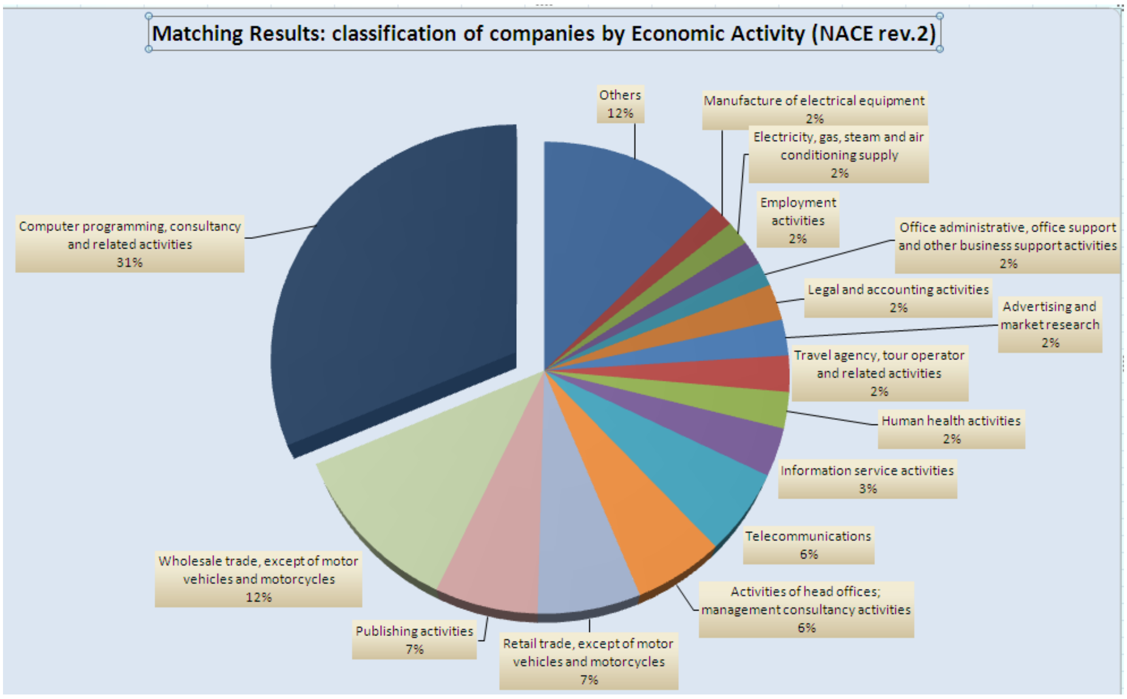
When matching to the JVS sample, our main interest was to explore to what extent the same company appeared in both the survey and dataset. It should be noticed that in our first experiment the matching results were very poor, in this second experiment, the results improved. 28% (34) of company names were matched. (See Figure 5). However, the response rate of these companies in the J-V survey was 38%. It was found out that there was no correlation between the number of on-line job ads and the results of the survey as regards J-V to be filled immediately.

Figure 5: Web Scraping 2^{nd} Experiment



The classification results of the companies by economic activity (NACE rev.2) reveal that 31% refer to computer programming, consultancy and related activity companies. However, there is a demand for IT-job from companies in all other Economic Activities in smaller percentages (Figure 6).

Figure 6: Classification of companies by Economic Activity (NACE rev. 2)

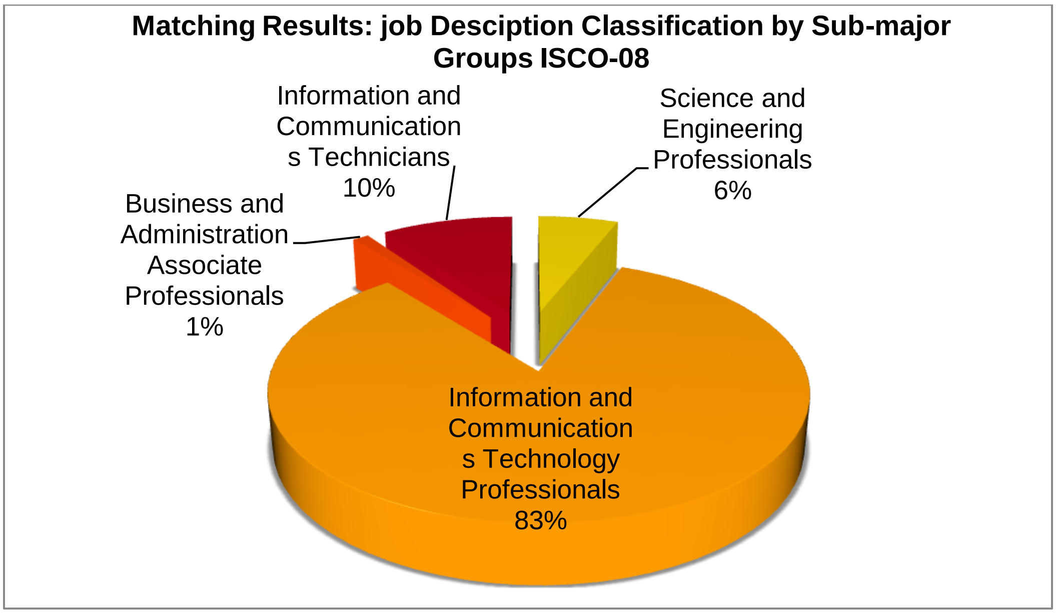


Given that the 2nd experiment concerned only the IT domain, we explored a little bit more the job description of the on line ads, since we have noticed that job title are not used in a consistent way (for example, an employee called a “Data Scientist” at one company has a different job title in another one). This is not peculiar due to the fact that there isn’t a clear definition for these new IT jobs. The results of the classification by 4-digit codes of ISCO-08 are shown in Table 2 and the results of the classification by Sub-major Groups ISCO-08 are shown in figure 7

Table 2: IT Domain – Percentage distribution of Job ads by 4-digit codes of ISCO-08

job_isco08	Title	%
2512	Software developers	30%
2513	Web and multimedia developers	14%
2514	Applications programmers	12%
2511	Systems analysts	10%
3512	Information and communications technology user support technicians	7%
2519	Software and applications developers and analysts not elsewhere classified	7%
2166	Graphic and multimedia designers	6%
2522	Systems administrators	6%
3511	Information and communications technology operations technicians	2%
2529	Database and network professionals not elsewhere classified	2%
2521	Database designers and administrators	2%
-	Others	2%
		100%

Figure 7: IT Domain - Percentage distribution Job ads by Sub-major Groups ISCO-08



Concepts and definitions of the basic variables used for the compilation of Job Vacancy Statistics pursuant to (EC) No 453/2008, (EC) No 1062/2008 and (EC) No 19/2009.

Job vacancy means a paid post that is newly created, unoccupied, or about to become vacant, 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 which the employer intends to fill either immediately or within a specific period of time. It should be noted that job vacancies refer only to employees.

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 is re-hired or returns to the enterprise after a holiday paid, or not, leave
- 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.

i) References

Anastasiou, S. (2014). Recruitment communication practices in job adverts in Greece through a snapshot of internet sites for job vacancies. International Journal of Economics and Management Sciences, Vol. 3, No. 2, p.p. 09-17

Galanaki, E. (2002). The decision to recruit online: a descriptive study. Career Development International, 7(4), 243-251.

Greece: Productive jobs for Greece. International Labour Office, Research Department. – Geneva: ILO, 2014

Moira, P., Milonopoulos, D., Anastasiou, S. (2004). Producing Graduates for the Tourism industry in Greece: a case study. Journal of Hospitality, Leisure, Sport and Tourism Education, 3(2), 55-60.

Terzis, V., Economides, A.A. (2005). Job Site Evaluation Framework (JSEF) and comparison among Greek and foreign job sites. Human Systems Management, Human Systems Management, Vol. 24, No 3, 223-237

Villar, E., Juan, J., Corominas, E., Capell, D. (2000). What kind of networking strategy advice should career counsellors offer university graduates searching for a job? British Journal of Guidance and Counselling, 28(3), 389-409.

Annex C: Slovenia

Introduction

The Statistical Office of the Republic of Slovenia (SURS) had previously gained experience in web scraping job vacancies data through the UNECE project which took place in 2014 and 2015. The main purpose of joining the ESSNet WP1 team was to continue this work and share the experiences among countries involved in ESSNet project. The objectives were:

- Investigate the population of job vacancy portals (and other specialized job agencies) in Slovenia and their characteristics
- Study the current methodology of creating statistics from current JV survey at SURS
- Establish a regular process of web scraping job vacancies data from the most important portals
- Process the scraped data (remove duplicates and link with business registers, etc.)
- Investigate the coverage of job vacancies derived from job portals and agencies in comparison of job vacancies disseminated from regular JVS.
- Brainstorm about possible experimental statistics

In 2016, the Slovenian team focused on identifying the main job portals and other agencies which advertise and establishment of regular collection (web scraping) of data from them. There was also established the process of editing the data collected from job sources.

At the end of 2016 the comparison between data from the regular survey and data from job portals, enterprise websites and Slovenian employment agency was made (Table 1).

Table 1: Comparison of job vacancy data sources, Slovenia

	Reported data	Scraped and administrative data	Job portals	Enterprise websites
Number of JV ads	4312	2321	1073	262
Percentage	100%	54%	25%	6%

Due to the fact the coverage from alternative sources was relatively low. Therefore the first task done in the first part of 2017 was ad hoc survey which was conducted along the regular job vacancy survey in February. The aim of survey was to collect the information about the channels which are usually used for Job vacancies ads (additionally the information about URLs of enterprises was sought). SURS Received about 3200 answers (out of 8800 units and survey) and first results show That around 20% of enterprises use social media (Facebook, Twitter, LinkedIn ...) as one of channels of advertising SE and almost 18% of them use other channels like print media, bulletin boards etc.

Detailed analyses of date will help SURS to incorporate this information in process of creating alternative JV statistics and testing the models of aggregation of scraped data.

In 2017 also we continue to collect the data from two major job portals every week. Additionally we also made a comparison of job vacancy survey data of first quarter 2017 with the combined data from Job portals, Enterprise websites and Employment agency of the Republic of Slovenia.

In April 2017 there was couple of meetings with the biggest job agencies which aim is to hire employees and outsource them to other companies or/and to conduct a recruitment of new employees in the name of other companies. SURS had a meeting with the biggest agency Trenkwalder International AG Personnel service provider (temporary staffing, permanent placement, HR services). Representative of Trenkwalder explained a process of recruitment new employees and channels of advertising. Due to the huge share of the all online Job Vacancies ads provided by those companies in Slovenia it is really important to distinguish the Job vacancies which aim is to employ the personnel at premises of specialized employment agencies and the recruitment of employees which will be employed in the company which hired agency for the recruitment process. SURS was also explained that it is not possible to distinguish these two types of employment from their job ads. One of solutions which were offered by them is to provide us list of job ads for which final aim is to find the employee which will be employed at the company who seeks for new personnel. In that case they would provide to NSI the information about activity of company, location of company and number of Job ads for each such company.

1. Methodology of the Job Vacancies Statistics at SURS since 2015

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 the target population. Included are business entities with registered main activity from B to S.

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), those 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.

Relevant JV: is every JV regardless of the day of advertisement which is unoccupied on the reference period

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 to replace. With this change we improve the quality of data and reduce the likelihood of double-counting the number of occupied posts.

Statistics: number of job vacancies, number of occupied posts and 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 sectors of activity: B-S

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 Centre at SURS, the reporting units contact them by phone or by email – the whole month after the reference day,
- by data collected from the Employment Service of Slovenia (for employers of the public sector and for state owned companies. They are still obliged to report JV to Employment Service of Slovenia).

Finally we combine the collected data with the 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, for seasonal adjustment of time series we use **JDemetra+** software, i.e. the TRAMO/SEATS method.

2. Job portals and specialised agencies in Slovenia

First task done was an investigation on existing job portals and other specialised agencies in Slovenia which advertise Job vacancies. We detected 9 job portals and 109 job agencies.

JOB PORTALS

General information

There are two main job portals (**Moje Delo, Moja Zaposlitev**) that advertise the biggest share of published job vacancies. The Slovenian team decided to collect (scrape) the data from those two job portals.

For the purpose of scraping, Agenty (Agenty.com) is used. APIs are 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 the next day. Job vacancies that were valid on the reference day are those, which were scraped on the first day and were published during that day. Additionally, if a job vacancy is being published more than a month, we treat it as invalid job vacancy.

On the day of scraping, each job portal is being scraped two ways: first we scrape a list of all job vacancies on the domain, and second we scrape a content of certain job vacancy. Following data are being scraped:

- 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)

Weekly about 2.000 – 2.400 queries are run (depends how many job vacancies are published).

Elimination of duplicates, which are the result of multiple scraping

As the web scraping is done weekly, some job ads 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 job ad page. Job ads for working abroad and for unpaid student work are also removed. This amounts to between 19% and 24 % of all job ads per week. Figure 1 represents the movement of all job vacancy ads scraped every week from mid-May to mid-September. The "cleaned" line represents data without ads for student work and work outside of the country. The trend is similar for both original and cleaned data sets and for both portals.

Figure 1: Weekly counts of all web scraped job vacancies before and after cleaning for selected portals

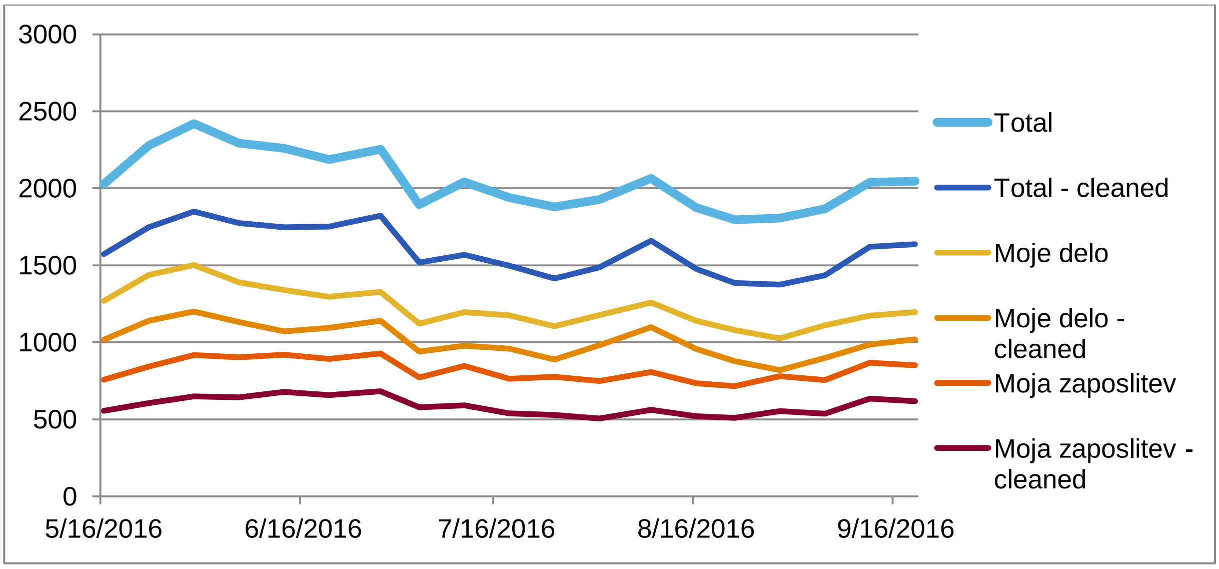
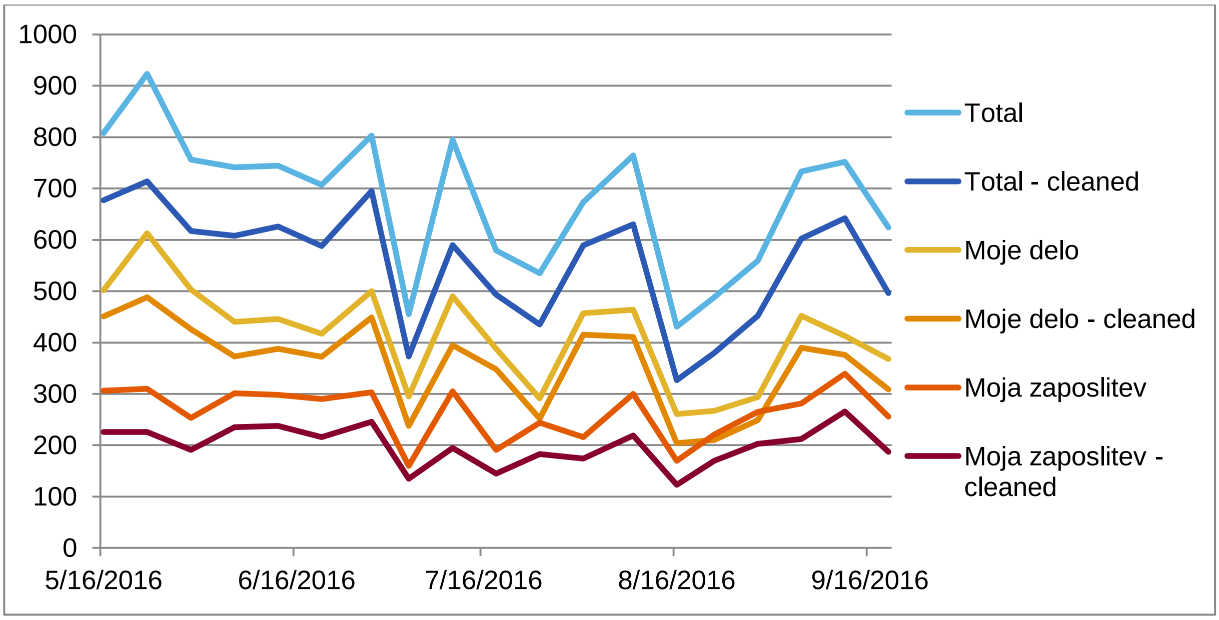


Figure 2. shows only new job vacancy ads and excludes those that have already been published. As on the first graph, trend movement is the same no matter if data set is cleaned of not and for both portals job portal, but the pattern is much more volatile.

Figure 2: Weekly counts of new web scraped job vacancies before and after cleaning for selected portals



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 match key for combining is company name. The aim of this task is to link registration number of a company to the job ad. We are able to do that for roughly 99 % of the job vacation ads.

The first step is to clean the data. All words must be written with capital letters, also the removal of some characters (e.g. .,&+-*/) is needed. Some companies do not have unique name and so we create two data sets: one with unique company names and one with duplicate company names.

We link 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 Register include a place of headquarters; if they do, we delete them and repeat the first way – direct comparison. 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, direct comparison produces duplicates. Therefore, when directly linking data with data set containing duplicate company names, we choose that observation where place of headquarters is the same as place of work. This tactic is only efficient when comparing to short name from the Register (step 4).

Additional 2.5 – 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.

We noticed that some companies publish a lot of job vacancies, but we are unable to link them by mentioned steps. Therefore we manually searched for proper registration number; a list of that kind of results has been made and is being updated after every record linkage (of new scraped data). This way we are able to link about 3 % with the "old" list and additionally we manually searched and found another 1.5 % of job vacancy ads. This is done in step 10.

The proportion of record linkage obtained via each step is similar for all time periods and the number of scraped ads (Table 2).

Table 2 Percentages of units linked with Business Register by method of linkage

Step	May 2016		Jun 2016		July 2016		August 2016	
	N	%	N	%	N	%	N	%
1.1	2109	73.28	2662	72.59	2192	73.95	2450	75.18
1.2	488	16.96	645	17.59	506	17.07	510	15.65
1.3	3	0.1	4	0.11	1	0.03	7	0.21
2.1	46	1.6	54	1.47	27	0.91	29	0.89
2.2	16	0.56	21	0.57	16	0.54	24	0.74
4	20	0.69	33	0.9	30	1.01	27	0.83
7	28	0.97	31	0.85	30	1.01	26	0.8
8	16	0.56	20	0.55	8	0.27	25	0.77
10	100	3.47	109	2.97	83	2.8	98	3.01
11	21	0.73	29	0.79	19	0.64	15	0.46
0	31	1.08	59	1.61	52	1.75	48	1.47
TOTAL	2878	100	3667	100	2964	100	3259	100

Removal of duplicates

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

Regarding duplication, we analyzed job vacancies ads data that were current on two reference days (31 May 2016 and 31 August 2016). The results are shown in Tables 3 to 5.

Table 3. Number of companies by job portals

Code	31.5.16		31.8.16	
	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 1 represents the number of all companies that published job vacancy ad on any job portal. 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 second largest portal. There were about 7% of companies that had job vacancies ad published on both portals.

Table 1.4 Number of companies by uniqueness of published ads

Code	31.5.16		31.8.16	
	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, that is 65%, 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%, publishes 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 positive, that they are duplicates (code 4).

Table 1.5 Number of job vacancy ads by job portals

Code	31.5.16		31.8.16	
	N	%	N	%
1	825	46.6	632	41.6
2	410	23.1	398	26.2
3	57	3.2	66	4.4
4	3	0.2	6	0.4
5	447	25.2	416	27.4
Total	1772		1518	

There were 40% of job vacancy ads 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, but there is not enough captaincy (code 4).

Code 5 represents job vacancy ads, published by job agencies. The amount of their job vacancy ads represents about a quarter of all job vacancy ads.

Employment agencies

Employment agencies have to satisfy a certain conditions, according to Slovenian law. They must be entered on to the Business Register with their main activity as *N78.200 - Temporary Employment Agency*. The Ministry of Labour, Family, Social Affairs and Equal Opportunities (MDDSZ) maintains the Register were such agencies are registered. MDDSZ issues a license to perform such work on territory of Slovenia. There are 109 employment agencies on the Register.

A quarter of agencies have less than 10 persons in employment. Some agencies don't provide temporary persons as their main activity, even though they are registered for such work. It was decided to survey the fifteen biggest agencies according to the number of person in employment and according to the number of ads for JV. In total are 13,700 persons in employment by agencies, 75% of them are employed by 15 agencies which we analyzed.

The two biggest job portals where companies can advertise the active job vacancies are MojeDelo and MojaZaposlitev. Companies can also advertise a JV via e-service by Employment Service of Slovenia (ESS) and public sector organizations are required to advertise via this portal. The ads which are advertised by employment agencies are problematic, as it will often not be known if the person will be employed by the agency or for the client. If it is the latter, then the client company is usually not mentioned in the ad.

We contacted 15 biggest agencies in order to find out:

- Where they usually advertise the JV (i.e. job portal, ESS)

- If they advertise JV also on their own web site
- If the advertised ad is for the agency (person will be employed by agency, but work for the client)
- Is the advertised ad for a client (the agency advertises, selects the appropriate candidate, but person will be employed by the client)
- Do they advertise for both (some ads are for the agency and some are for the client)

Results:

Only 33% of agencies advertise JV exclusive for themselves but only have 6% of JVs (Table 1.5). In this case data obtained by web scraping data from job portals will be correct. However, the majority (67%) of employment agencies advertise both for themselves and for client. Agencies of this type advertise the majority of JVs (more than 94%). As we don't know whether the employer is the agency or the client, then some of the on-line job ads will be incorrect (although we won't know which ones).

Table 6: Aggregated data for 15 agencies by advertised ads, September 2016

	Number of agencies	Share of agencies	Number of JV	Share of JV
advertise exclusively for themselves	5	33%	19	6%
advertise exclusively for the client	0	0%	0	0%
combination of ads (some for agencies and some for subscriber)	10	67%	320	94%
Total	**15**	**100%**	**339**	**100%**

Future plans involve contacting agencies and invite them to a meeting to explore the possibility of getting more information. Scraped data from job portals can't be used without additional information which only employment agencies have.

Below in table 7 there is more detailed data by each agency.

Table 7: Detailed data by agencies

	Agency	ID_Number	URL Address	Persons in employment (size class)	The agency advertise JV for themselves (**person is employed by the agency)**	The agency advertise JV for subscriber **(person is employed by other company)**	In which job portal the agency usually advertise JV? (1=MojeDelo, 2=MojaZaposlitev, 3=Employment Service of the RS)	Do they also advertise JV on their home page?	Number of active ads for JV (at the end of September)
1	**ADECCO H.R. d.o.o.**	1519140000	http://www.adecco.si	>900	Yes	Yes	1	Yes	87
2	**KARIERA D.O.O.**	2111357000	http://www.kariera.si	>900	Yes	Yes	1,2	Yes	46
3	**TRENKWALDER D.O.O.**	1622226000	https://www.trenkwalder.com/si	>900	Yes	Yes	1, 2	Yes	64
4	**ISFACILITY SERVICES D.O.O.**	5472784000	http://www.si.issworld.com	>900	**Yes**	**No**	2, 3 (exceptionally)	Yes (link to MojeDelo)	**5**
5	**NATON D.O.O.**	1889664000	http://www.natonhr.com	500-899	Yes (90%)	Yes (10%)	1	Yes	9
6	**AGENCIJA M SERVIS D.O.O.**	2160544000	http://www.mservis.si	500-899	Yes	Yes	1	Yes	32
7	**PAPIR SERVIS D.O.O.**	5226236000	http://www.papir-servis.si	500-899	**Yes**	**No**	3	No	**-**
8	**MANPOWER D.O.O.**	5674115000	-	100-499	Yes (they have less ads for themselves)	Yes (they have more ads for clients)	1, 2	No	34
9	**KOROTAJ D.O.O.**	2091933000	http://www.korotaj.si	100-499	**Yes**	**No**	1	No	**0**
10	**KI INTERIM D.O.O.**	2156199000	http://www.interim.si	100-499	Yes	Yes	1,2	Yes	8
11	**ATAMA D.O.O.**	1318527000	http://atama.si	100-499	Yes	Yes	1, 2	Yes	15
12	**FMG KADROVSKA AGENCIJA D.O.O.**	3698530000	http://www.fmg.si	100-499	**Yes**	**No**	1	Yes (link to MojeDelo)	**10**
13	**AXENT D.O.O.**	1254669000	http://www.axent.si	100-499	Yes (1%)	Yes (99 %)	1, 2	Yes	8
14	**KAČAR-MONT D.O.O.**	5611393000	http://kacar-mont.si	100-499	**Yes**	**No**	2, 3 (exceptionally)	No	**4**
15	**POWERSERV D.O.O.**	2139430000	http://www.powerserv.si	100-499	Yes (90 %)	Yes (10 %)	1, 2	Yes	17
	Total persons in paid employment		-	**10,319**					**339**

3. Results of ad-hoc survey about the modes of JV advertisements

After the reference period of regular survey which was the first quarter of 2017 (28th February), the evaluation questionnaire was sent to units selected in the current sample with the aim of collecting information about the modes of their advertisement of job vacancies. The main question in the questionnaire was:

“Do you usually advertise the JV ads directly or you use an employment agency?”

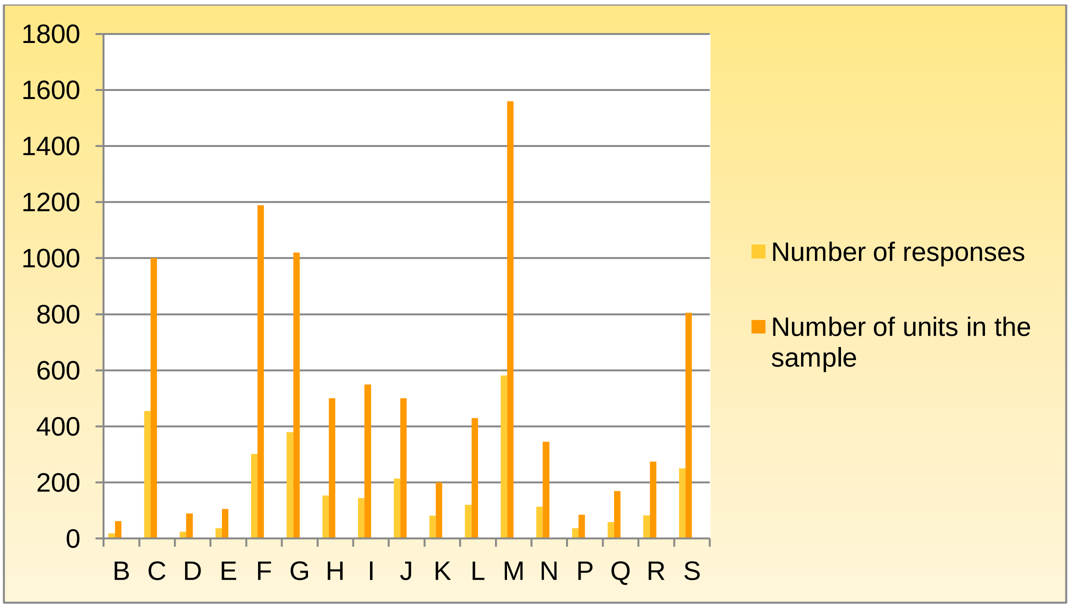
For those units which answered “directly” the subsequent question was:

“What channels you usually use for advertising job vacancies?” Multiple answers were possible:

- Job portals
- Employment agency of the Republic of Slovenia
- Agencies (e.g. Manpower, Adecco, Trenkwalder …)
- Own website
- Facebook
- Other social media websites (LinkedIn, Twitter…)
- In the press
- Other

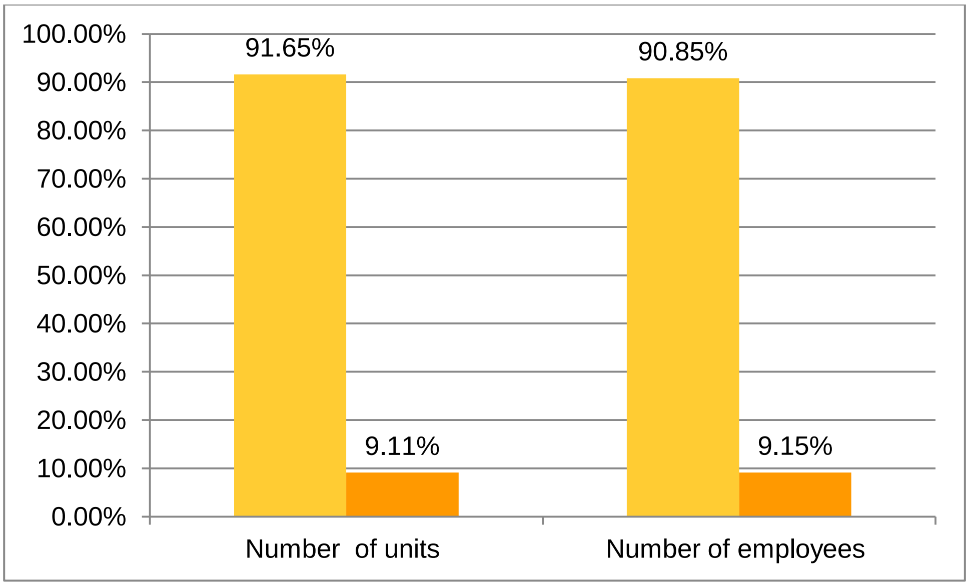
Out of 8885 units in the sample SURS received 3054 valid answers. Due to high non-response, we are not able to say how representative are our set of responses. However, we can say that the distribution of responses broken down by sectors of activity is similar to the distribution of all units from the sample (Table 7).

Table 7: Number of units in the sample and number of responses broken down by activity



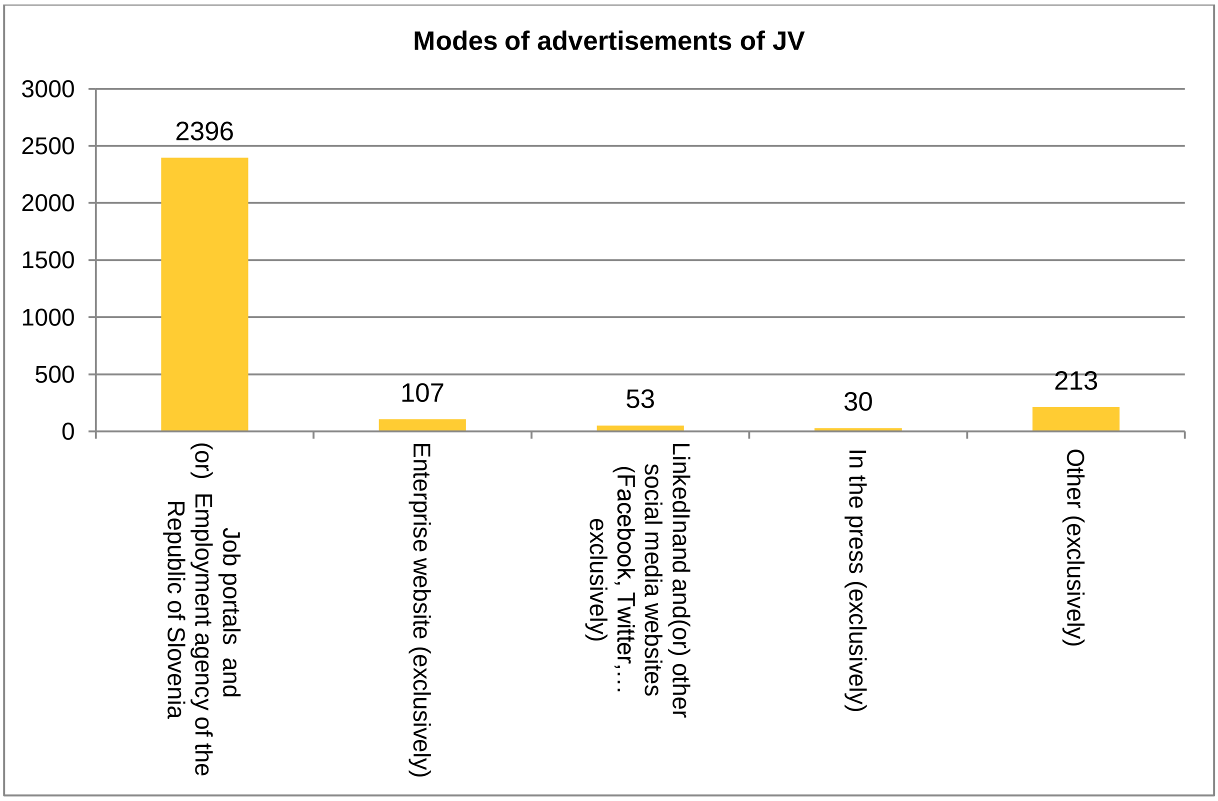
Of those reporting units that responded, over 90% advertised jobs directly, with less than 10% using employment agencies (Table 3.2).

Figure 3: Distribution of units that advertise JVs directly and via agencies



If we attach reported data about advertised JVs from the regular JV survey, around 95% of reported data come from units which advertise JVs through job portals and (or) employment agencies and just around 5 per cent of reported data come from other channels.

Figure 4: JV advertising channels



4. Time lag between the advertisement of job vacancy and actually employment

The timing of a job advertisement does not correspond exactly to the definition of a vacancy in the JV survey. In particular, a vacancy still exists from when an advertisement closes until it is filled by the employer. In order to determine which ads are valid in the certain period it is important to understand this time lag.

This can be estimated using data from the Health Insurance Institute of Slovenia. Every enterprise is obliged to provide the information about the social insurance of employees to the Health Insurance Institute. This is provided using the M-1 form (Reports of data regarding the pension and disability and health insurance, parental insurance and unemployment insurance). The M-1 form contains many variables including the enterprise ID, the starting day of insurance and job advertisement code, which allow us to link the M1 data with the record from the Slovenian Employment Agency where the JV is advertised.

For the purpose of calculating the average distance between the starting day of advertisement of a certain JV and starting working day monthly M-1 data of January, March and April of 2013 is investigated. Each record from M1 data was linked by code of advertisement to records of Slovenian Employment Agency from January 2012 onwards. The below table shows average time legs for each observed month and average time lag in first quarter of 2013. There is on average about a 45 day lag from when a job is first advertised to when it is actually filled.

Table 8: Time lag between advertising a job vacancy and filling it

Time period	Average time lag (days)
January 2013	44
February 2013	46
March 2013	45
First quarter of 2013	45

5. Legal and ethical issues of collecting the data from Job portals

Exploring the legal and ethical issues SURS found out that scraping JV from job portals is more ethical then legal issue. There are no legal acts currently which will prevent one to scrap data from websites. However from the ethical point of view it would be not ethical to produce official statistics out of scraped data where owner of website disagree with collecting of its data. Moreover owners could always block the process of scraping the data from the websites so it is crucial to talk with them and tray to establish the partnership.

It was also found out that majority of websites contain a sub link “\Robots.txt” where the information of which contents of websites are allowed to be collected by robots. This is especially important for Job portals. It is expected from one who wants to scrap the data from such website to follow the rules which were stated in the text document of robot.txt and consequently minimizes burden to owners.

The description of in the “robot.txt” usually contains following items:

- User-agent: contains names of robots which are not allowed
- Disallow: contain the list of directories for which “web-scraping” is not allowed
- Allow: contains exceptions (datasets) in the above directories for which “web-scraping” is allowed
- Crawl-delay: The idea is that the “web-scraper” uses a few seconds delay between successive crawling per page.

6. Conclusions and future perspectives

During the period of SGA1 ESSNet Big Data project Statistical Office of the Republic of Slovenia has made good progress in investigating the challenges related to all potentials of JV data collected from job portals and employment agencies.

The main goals achieved are:

- Regular collecting data about job vacancies from the main job portals in Slovenia has been established.
- Collected data is processed by statistical methods of cleaning and linking with the Business registers.
- Channels of advertising job vacancies were investigated.
- Ethical and legal issues related to web scraping of JV were investigated.
- The methodology of collecting and processing JV data was developed.

We had some difficulties with understanding the big difference of coverage of JVs from job portals and actual JVs from the survey. The main reason is the lag between when jobs are advertised and when they are filled. This information will allow us to better understand the differences between sources. Regarding legal and ethical issues SURS made significant progress in negotiating with owners of main job portals in Slovenia. We are just about to sign the contract with owners of portals which will allow us officially scraping the JV data from job portals.

SURS fulfilled many of the preconditions needed to focus on preparing a possible set of experimental statistics, which will be our primary focus for the SGA2 period.

Annex D: Sweden

Introduction

Since the beginning of this year, our efforts have focused on acquiring data from private job portals. We are in contact with three private job portals, chosen from twelve portals previously identified as the most important in Sweden (Deliverable 1.1).

Two of the three portals provided us with data; Metrojobb.se and Jobbsafari.se. Both are hybrid job portals, i.e., they extend their own job advertisement boards with the results from search engines. Metrojobb provides us with their own advertisements only, while Jobbsafari provides the complete set of data at a given point in time. The third company, Textkernel, does not collect any data for Sweden, but they gave us valuable information about their methods. They are prominent in de-duplication, profession identification by the advertisement job titles, and data source tracking. They use semantic search and machine learning methods.

Besides the two portals, we have since earlier in the project data from the job portal of the Employment Agency. As expected, their data have the best quality and the most complete description of opening jobs compared to the other two portals. The Employment Agency estimates that their portal covers ~40% of the online job advertisement market.

We started by evaluating the data from each portal separately. The data from the Employment Agency covers the period January 2012 to May 2017. From Metrodata, we started collecting data at the beginning of March 2017, and from Jobbsafari at the beginning of April 2017.

We thus now have data from all three portals for the period April to May 2017, and data collection continues. The description below focuses on data from April 4 to May 17, 2017. For this period, data were de-duplicated and the three sources were integrated.

The preliminary results show that the quality of the data varies a lot between the different portals, and their data can be useful for different purposes. With the Employment Agency data, we tested algorithms for identifying professions (using text mining). We can link data to the Business Register since unique organization identification is available in the data, and we can link to the sample used in the Job Vacancy Survey, at the local unit level.

Jobbsafari has the ambition to detect as many advertisements on Internet in real time as possible. Our hypothesis is that their data has the highest coverage of online advertisements of jobs, but the quality of the information is lower.

Metrojobb has a large portion of advertisements from outsourcing companies and many advertisements stay online for quite a long period. We may learn features of this type of advertisements since identification of so-called ghost advertisement is an issue we will have to solve.

Access to data

Since the legal issues for web scraping for official statistics are unclear in Sweden, we were not able to take a general web scraping approach for data collection. We contacted job portals and reached agreements in the common interest of better official job vacancy statistics. Some lessons learned so far:

- Even with an agreement, it is essential to behave properly and follow ethical standards for the web scraping, and for example make it simple for the portal owners by using their available APIs.
- Close communication is essential in order to confirm the interpretation of variables. If confirmation is not possible, we have to understand the data by data exploration. It would be desirable to develop common methods for standardization of variables.
- In our experience, the contacts and discussion have been very open and there have been no difficulties to reach an agreement on data for testing purposes. We find that our approach could be considered by NSIs as an alternative to investing in a web scraping system. Web scarping is complicated and expensive; however, an in house system may be a more stable data source. The private companies come and go, while the Internet remains as the communication channel and platform.

Python is our main processing tool for API data retrieval, data exploration and uploading to the SQL server. From the Employment Agency, we received xml files in bashes. These files are processed with python scripts. The libraries used are xml.etree.ElementTree, pandas, sqlalchemy, and re.

From Metrojobb.se, we download data with APIs once a week. The interval was decided by analyzing the data of the Employment Agency. The mean period of an advertisement existing online in the database is 26 days, and the median is 21 days. A one-week interval is likely to catch the total amount of advertisements. The data is in json format, which is a common format used for Internet communication. The python scripts use libraries urllib, json, pandas, hashlib, and BeautifulSoup.

From Jobbsafari.se, we download data daily via API. Jobbsafari suggested the downloading interval. The downloaded format is in xml. The python scripts use libraries such as sqlalchemy, re, hashlib and pandas.

We use SQL server for the storage of the data and some basic analysis. SAS or R will be used for further analysis.

Description of available data

The aim at this point was to integrate data from the three available job portals, concentrating on the overlapping period, i.e. all job advertisements that have a starting date between April 4 and May 17 2017. Table 1 gives an overview of available variables and the rate of valid values (the percentage of correctly spelled, sound values and non-missing values).

Table 7 Data overview

Variables	*Rate of valid values, %*		
	Swedish Employment Agency	*Metrojobb*	*Jobbsafari*
Name of employer	100	100	-
Title of advertisement	100	100	100
Description of advertisement	~100	~99	74*
Starting date	100	100	100
Ending date	100	100	-
Scraping date	-	-	100
Publishing date	-	100	-
Data source	-	100	100
Number of jobs in advertisements	100	-	-
Municipality of work place	99	~98	~87
County	100	~100	-
Code for profession, high level	100	~100	-
Code for profession, intermediate level	100	-	100
Code for profession, detailed level	100	~48	-
Organization id	~100	-	-

*it is snippet of description, not like other sources with complete description.

The Employment Agency provides the most complete data in terms of available variables and full text of the advertisement, and the rate of valid values. The total number of unique advertisements collected for the period is 101 316.

In total there are 22 994 employers presented in the Employment Agency data. The top six employers are all outsourcing companies, and their share of the advertisements is 5.5%.

All the 290 municipalities in Sweden are represented.

During Sprint 2, we made a first attempt to link to the Business Register, and since then our method has been much improved. Almost all advertisements can be linked to the Business Register by organization id. This identification is at the level of enterprise. For linking to the local unit level, we used the address. The preliminary result is that 53% of the addresses can be matched to the Business Register. With the connection to the Business Register, it is possible to add a lot of useful information such as profession and branch.

As a small experiment, we tested text analysis on a sample of advertisements from the Employment Agency data. The idea was to evaluate the possibility to classify advertisements according to job descriptions, using the fact that we have information about profession from the Business Register. First results are promising and we will continue this work during SGA 2.

The Employment Agency data includes the variable Number of job in advertisement. This is a useful variable since one advertisement may advertise more than one vacancy. This is the case for 22% of the advertisements in the Employment Agency data.

Metrojobb have relative good quality of their data. The total number of unique advertisements collected for the period is 56 693. Few values are missing or of bad quality by appearance.

Similar to the Employment Agency data, the employers with the largest share of the advertisements are consulting and outsourcing companies with one exception (Stockholms kommun).

There are 288 municipalities represented in the dataset.

Metrojobb have two codes for professions, at a high and at a detailed level. They are internal and do not follow any standard. We have not yet tried to translate them into the standard code.

The starting and ending dates refer to the advertisement of the vacancy, while publishing date is the date when the advertisement appears on line. Publishing date may differ from the starting date.

Jobbsafari had in total 145 923 unique advertisements during the period, including advertisements scraped from the Employment Agency and Metrojobb.

Since Jobbsafari makes great efforts on detecting new job advertisements online, it is interesting to look at the sources. The sources are the urls from which the advertisements were scraped (variable Data source). Only ~5% are unique for Jobbsafari. 60% are from the Employment Agency and 12% are from Metrojobb. In total, there are 149 sources. This information is useful for determining coverage on the Internet.

The quality of the municipality information is not high and it does not follow the standard for Swedish municipalities.

The starting date and the scraping date are identical in the dataset, which indicates that Jobbsafari identify new ads in real time, i.e. they detected all advertisements at the same day as they started.

The code on profession does not follow any standard and can only be used as a reference.

Data handling

Before merging data, we cleaned each data set for duplicates.

For Metrojobb, we used the variables Name of employer, Title of advertisement, Description of advertisement, Municipality of work place, and the starting and ending dates. Before de-duplication, there were about 91 000 advertisements, and about 38% were duplicates.

In total, 159 132 advertisements were collected from Jobbsafari. The variables Title of advertisement, Description of advertisement, Municipality of work place, and the starting date were used to identify duplicates. There were 8% duplicates in the dataset.

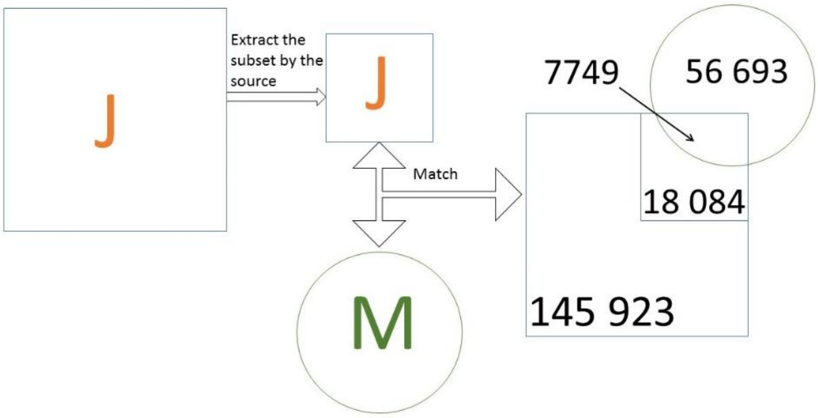
Although the Employment Agency checks for duplicates before they deliver data, we identified duplicates with the variables Name of employer, Title of advertisement, Description of advertisement, Municipality of work place, and the starting date. About 4% duplicates were found.

For merging the datasets, we start from the advertisements from Jobbsafari. They are first integrated with the Metrojobb data, and then with the Employment Agency data.

For integration, two strategies were tested. The first strategy used the variable Data source provided by Jobbsafari, which is the link to where data was scraped. For the advertisements scraped from Metrojobb, the source is of the form "http://metrojobb.se.....".

We extracted the subset of data from Jobbsafari with source Metrojobb, 18 084 advertisements, and this subset was then matched with Metrojobb by the variables Title of advertisement, Municipality of work place, and the starting date. Only 7 749 advertisements were found to match, see Figure 1.

Figure 1 Integration of Jobbsafari and Metrojobb data



Many advertisements published on Metrojobb did not have matches in Jobbsafari. It could be the case that Jobbsafari scraped the same data from several sources and deleted the ones from Metrojobb as duplicates. We know that both Jobbsafari and Metrojobb do take away duplicates from their datasets, but not how. However, it is not satisfactory that about 59% of the advertisements scraped from Metrojobb and published by Jobbsafari were not possible to match to Metrojobb. The variable Municipality of work place is not standardized in Jobbsafari; these can be postcode or non-municipality names. The standardization of variables may change the matching figures.

Using a second strategy of integration, we compared directly Jobbsafari and Metrojobb by the variables Title of advertisement, Municipality of work place, and the starting date, ignoring the variable describing sources in the Jobbsafari data. 15 913 advertisements were identified as duplicates. From these duplicates, we can see that some of the Metrojobb advertisements also exist in other sources, even though Metrojobb claim that they deliver only the jobs posted only on their own job portal. Obviously, many advertisements are posted on multiple channels. We found this strategy better, since multiple channels can be identified.

The situation is similar when integrating Jobbsafari with the Employment Agency. With the second strategy, 43 550 advertisements are matched when we combine Title of advertisement, Municipality of work place, and the starting date.

Of around 304 000 advertisements from the three portals, about 6% duplicates were identified, leaving 287 000 unique advertisement in the merged data set.

Methodology

A first attempt of quality assessment of the data from the Employment Agency was made during the third virtual sprint of SGA 1 in February 2017. We continue with this work and we are currently investigating the possibility to assess the quality of the merged data set. We use two approaches; A framework originally intended for administrative data sources and applied by the statistical office of New Zealand (Zhang 2012[13], Reid et al 2017[14]), and a framework suggested for big data sources suggested by the UNECE Big Data Quality Task Team (UNECE 2014[15]).

The UNECE framework has three "hyper dimensions"; the source, the metadata, and the data, with quality dimensions nested within the hyper dimensions. These dimensions apply to the three phases of the business process: input, throughput, and output. The framework for integrating data proposed by Zhang (2012) could further deepen the Accuracy and Selectivity dimension of the UNECE framework by a total survey error approach. In particular, the coverage of the on line sources needs to be established, if possible.

Future work

We will continue the quality assessment of the merged data set. A fully developed framework is useful when defining models for estimating relevant statistics, and comparing those with survey estimates. Data from the Job Vacancy Survey are available for the same period as the Employment Agency data.

[13] http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2011.00508.x/full

[14] https://www.degruyter.com/downloadpdf/j/jos.2017.33.issue-2/jos-2017-0023/jos-2017-0023.pdf

[15] A Suggested Framework for the Quality of Big Data - Deliverables of the UNECE Big Data Quality Task Team, December, 2014

We intend to make more use of the Business Register and further develop the text analysis described above. A categorization algorithm could help us improve the quality of the data from the other portals. Another example where algorithms could help improve data quality is to correct for spelling errors and thereby improve matching and de-duplication.

We will continue to consider additional sources. Statistics Sweden has recently signed an agreement with Vainu, a company that do web scraping of general business information. We will investigate whether the services they offer can be helpful for collecting job advertisements.

Annex E: United Kingdom

1. Introduction

During 2016 the UK pilot was focused on exploring different web scraping approaches for different job portals, cleaning data, identification of duplicates, matching to survey data and the development of a high level conceptual framework for using on-line job advertisements for statistical purposes. These are documented in the SGA-1 Interim Technical Report. Although we learnt a lot, it was clear that there are a huge number of challenges and that we do not have the resource to tackle all of them. We decided then to focus on tasks that would add the most value.

For this reason, since the beginning of 2017, the UK has focused on developing a framework for web scraping counts of vacancies by company name from different job portals and then matching these to reporting units in the job vacancy survey (JVS). This provides a basis for directly comparing vacancy counts by company from different on-line sources to the JVS and therefore understanding the differences between them. In addition, a framework was developed for scraping vacancy counts from enterprise websites. The initial intention of the pilot was that web scraping of enterprise websites would be done during SGA-2. However, we made an assumption that enterprise websites vacancy counts are the "gold standard" of on-line sources and so it would be useful to see how job portals would compare against them.

This approach tackles one of the most important issues with this type of data, namely the representativeness of on-line job advertisements compared with the target measure of the JVS. This framework allows us to identify the specific companies and industries where there are large differences between what is advertised on-line and what is reported via the JVS. The framework can also be applied for different job portals and then used to identify which portals have the best coverage. Finally, this also provides a framework for integrating these data and so moves us towards being able to use these data for producing statistical outputs. The current framework is a prototype that collects a very narrow amount of information (i.e. counts of vacancies by company) and has only been developed for 50 companies. However, this is an approach that could be easily scaled up.

The structure of this report follows the general approach set out for the ESSNet as a whole, namely:

- Data Access
- Data Handling
- Methodology
- Statistical Outputs
- Future Perspectives

The main emphasis of this report is on the most recent developments since the beginning of 2017, although some aspects from the earlier investigations are included for completeness.

2. Data Access

As discussed in the first Big Data ESSNet WP1 Deliverable[16], there are a large number of job portals in the UK and the overall data landscape is very complex. The UK pilot has explored both direct web scraping and arranged access.

2.1 Direct web scraping:

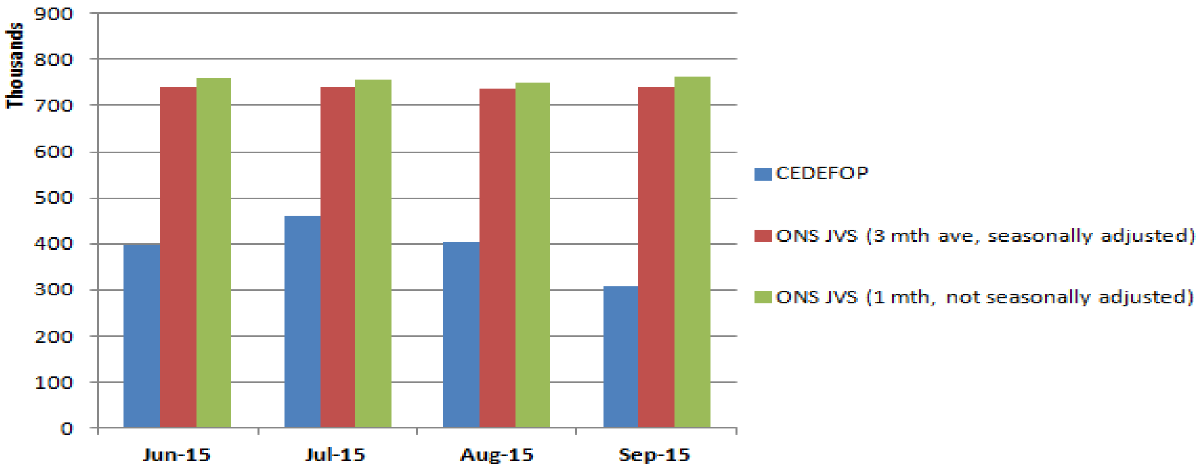
ONS is still developing its policy on web scraping but the main working principle is that web scraping activities should respect the terms and conditions of target websites. If a website prohibits web scraping then explicit permission or alternative arrangements should be agreed with the website owner. However, there can be some latitude when undertaking small scale experiments as the legal, reputational and ethical risks are negligible. The results of these small scale tests provide the basis for identifying those websites that could offer the most benefit and therefore, where future engagement activities should be focused.

2.2 Arranged Access:

2.2.1 CEDEFOP

The UK was one of five countries included in the 2015 CEDEFOP pilot. The UK data was sourced from four job portals (Monster, CV-library, Reed, and Totaljobs). Various steps were taken to clean the data, remove duplicate job ads and classify variables. An assessment[17] was made by the Marchmont Observatory, which provided UK support to the project. This found that monthly volume of job adverts captured were on average just over half that of the official estimates based on the ONS JVS. There was also much more variability in the monthly CEDEFOP figures, compared with the ONS JVS (Figure 1).

Figure 1 Number of job vacancies, CEDEFOP pilot versus ONS estimates



Source: Marchmont Observatory, ONS

[16] Inventory and Qualitative Assessment of Job Portals: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/File:Deliverable_1_1_final.docx

[17] CEDEFOP UK County Report, Marchmont Observatory – University of Exeter (unpublished)

The official ONS job vacancy estimates are seasonally adjusted based on a three month average and the UK pilot these are not directly comparable with the CEDEFOP figures. However, even the non-seasonally adjusted monthly figures also show far less variability than the CEDEFOP figures suggesting that there are more fundamental issues with the data. The Marchmont Observatory study identified definitional differences between sources as a likely issue. The JVS measures the stock of existing vacancies, where as the CEDEFOP data apparently measures all new vacancies collected during the reference period.

2.2.1 Other third-party providers

The UK pilot has explored other avenues for gaining access to data with third parties. Preliminary discussions have been held with several job portal owners where the website terms and conditions around web scraping were restrictive and/or API access was limited. Initial discussion indicated that some companies might only agree to supply data on the condition of some form of payment. Given a possible requirement to purchase data, ONS Procurement has become involved. They advised that all potential data suppliers of on-line job advertisement data should be given an equal opportunity to partner with ONS. On the 1 June 2017, a market engagement notice was posted on the ONS tender notification system inviting owners of on-line job advertisements to discuss opportunities for partnership with ONS[18]. This process does not involve any promise of direct payment. As of the end of June 2017, three organisations had registered an interest.

2.3 Summary:

There are a number of possible routes within the UK for acquiring on-line job vacancy data for statistical purposes. However, there are legal issues still to be resolved before being able to web scrape data at scale. Steps are being taken to resolve these issues by looking to form data partnerships. This approach may also give us some much better access to data that may also be of better quality and so this will be a key focus in the next phase of work.

3. Data Handling

3.1 Web scraping approaches used during 2016:

Our early web scraping approaches focused on point and click web scraping of job portals (using Import.io) and extracting data from job portals using public APIs, specifically those provided by Adzuna, Indeed, and Universal Jobmatch. Our conclusion was that while Import.io was useful for small scale, one-off web scraping experiments it had more limitations than more programmatic approaches (e.g. Python Scrapy). In terms of APIs, while we found these useful for collecting a lot of data, they also had limitations. The Universal Jobmatch API had a limit on the number of job ads that could be returned on any search criteria. Those from Adzuna or Indeed could not be used to query by company name. None of the APIs provided the full job description – only a "snippet" of a certain

[18] https://in-tendhost.co.uk/ons/aspx/ProjectManage/263

number of characters. For these reasons, we have started to focus on more programmatic approaches to web scraping.

3.2 Newly developed web scraping approaches:

The web scraping framework developed during 2017 has focused on matching counts of job vacancy by company to reporting units in the JVS. This has involved taking a sample of 50 reporting units out of about 1300 that always included in the sample. In total, three distinct web scraping approaches were explored for obtaining on-line vacancy counts:

i) Portal job vacancy counts (PJVC) scrapers:

The purpose of these scrapers is to collect vacancy counts for the selected reporting units from selected job portals. They work by running through a manually constructed list of URL suffixes - one for each reporting unit from the sample - pointing to a page giving the portal's job vacancy count for that given reporting unit.

For example, the URL *https://www.indeed.co.uk**/HSBC-jobs*** returns a webpage containing Indeed's current JV count for HSBC and so the URL suffix is: ***/HSBC-jobs.***

For Adzuna the equivalent URL is: *https://www.adzuna.co.uk**/jobs/company/hsbc*** and so the URL suffix is: ***/jobs/company/hsbc***

For the most part, the list of URL suffixes has to be created separately for each portal. However, once one list has been created there are ways of speeding up others. For example, an existing suffix list from one portal may be reused to find some valid page hits, with the residuals being searched manually. Each portal has a different format and structure and so a separate robot is needed for each one.

PJVC robots were developed for the following job portals:

- Indeed
- Adzuna
- Careerjet
- Brick7
- Jobijoba

The robots run daily with and the resulting output from each run looks something like this:

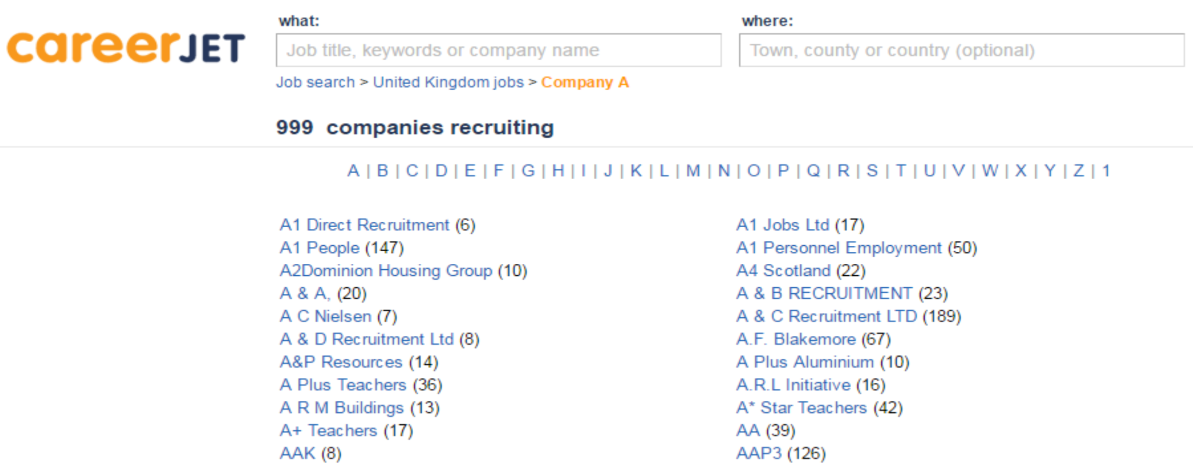
Table 1: Sample output from PJVC scraping:

JV Count	Indeed	Adzuna	Careerjet	…..
Reporting Unit A	105	121	78	
Reporting Unit B	56	56	17	
Reporting Unit C	505	520	532	
Reporting Unit D	12	15	12	
….				

ii) <u>Portal Company Page (PCP) scrapers</u>:

Several of the large UK job portals have a company directory. A screenshot from Careerjet is shown in Figure 2. In this case, the company names and vacancy counts can be scraped directly from the company directory URL. This is a much easier and quicker approach for collecting company counts. However, the challenge is then to match and reconcile the scraped company names back to the JVS reporting unit. This is covered further in Section 4.2.

Figure 2: Careerjet company directory screenshot



careerJET

what: Job title, keywords or company name

where: Town, county or country (optional)

Job search > United Kingdom jobs > Company A

999 companies recruiting

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | 1

A1 Direct Recruitment (6)
A1 People (147)
A2Dominion Housing Group (10)
A & A, (20)
A C Nielsen (7)
A & D Recruitment Ltd (8)
A&P Resources (14)
A Plus Teachers (36)
A R M Buildings (13)
A+ Teachers (17)
AAK (8)

A1 Jobs Ltd (17)
A1 Personnel Employment (50)
A4 Scotland (22)
A & B RECRUITMENT (23)
A & C Recruitment LTD (189)
A.F. Blakemore (67)
A Plus Aluminium (10)
A.R.L Initiative (16)
A* Star Teachers (42)
AA (39)
AAP3 (126)

PCP scrapers were developed for Universal Job Match, CV-Library and Careerjet. Universal Job Match lists the 200 largest companies and it was later established that many of these companies were employment agencies or other job portals (e.g. Adzuna) and so this directory was not considered useful. CV-Library had the largest company directory with almost 10,000 companies. Unfortunately, most had much lower counts (often nil) compared with other on-line data sources. It is assumed that this is because it is a job board and so only advertises jobs that have been uploaded by the employer. However, the CV-Library company directory may still be very useful as it contains a URL to the careers page of each company in the list. These could be used as input for an approach to scrape company web pages directly. Although the Careerjet directory only had about 1000 companies, the vacancy counts were more comprehensive. This is presumably because Careerjet is a job search engine and is not reliant on jobs being uploaded by employers.

It is worth noting that although other portals had some kind of company directory, these did not always produce reliable results. For example, the entry "Brown" in the Brick7 company directory, retrieved jobs for "Eden Brown", "Foxwell Brown" and "Currie and Brown UK Ltd". Indeed had company profile pages containing job counts, but no company directory page from which these could be accessed.

iii) Company Web Page (CW) scrapers:

The third web scraping approach was obtaining job vacancy counts directly from the web site corresponding to the survey reporting unit. This involved manually searching for and identifying the company web page of the enterprise, identifying the jobs section of each website and then developing scrapers to extract counts of the number of current job vacancies. This has the same objectives as the approached being developed in WP2 (Web scraping for enterprise statistics) except that it is less automated.

The first step was to manually inspect both the terms and conditions and the robots.txt exclusion protocols of the 50 reporting units. Websites could be found for all these reporting units. However, four did not allow web scraping while one did not advertise any job vacancies of their website leaving 45 for which a count of current job vacancies could be scraped. The amount of work involved in manually writing 45 robots using this approach was initially quite labour intensive. In an effort to speed this up, a framework was developed to extract all the commonalities across different robots. Some existing code could then be reused and only code specific to the company website needed to be written. This framework uses elements from Python Scrapy.

The method for extracting the vacancy count depends on the design of the target website and how information about job vacancies is presented:

- Some websites contain a specific value that represents the count of current job vacancies (e.g. "Job Openings 1 - 25 of 37"). In this example, the value “37” can be extracted using a simple HTTP request followed by XPATH and regex to extract the target value.
- If this value does not exist, then the count is derived by summing common HTML elements corresponding to each individual job ad.
- Some company webpages do not have with a large number of jobs use pagination and so a loop is needed to load all the job ads.
- Some webpages load the desired content asynchronously, thus it is not present in the initial HTTP response. In those cases Selenium was used, which simulates visiting the URL with a web browser (via PhantomJS).

Of the 45 enterprise spiders developed:

- 25 used XPATH + regex, 20 used a job ad count method
- 4 required a pagination loop, 41 did not.
- 31 used a simple HTTP request and 14 used Selenium

Having developed this framework we estimate that 20-25 robots a day could be developed using this approach. It might be possible to speed this up further by providing some kind of interface to create the robots by assembling prefabricated code depending on the design elements of the website. This is more of a “brute force” method than the generic web scraping approaches being developed in WP2. However, if one considers that there are 1300 large reporting units that are always in the UK JVS, we estimate that it would take no more than a few person months to develop robots for all of them.

3.3 Web scraping IT infrastructure

All robots are written in Python. The PJVC scrapers all simple HTTP requests, while PCP and CW spiders use Python Scrapy. All scrapers introduce a delay between the requests and refrain from parallel requests so as to minimize the load on the target websites. After scraping, the robots store the data in Mongo DB hosted in Google Cloud. The robots are deployed using a custom-made deploy script to a Google Compute virtual machine instance, where they are run every night by a Cron scheduler. For the most part, these technology choices have been informed by solutions already developed for other prototype web scraping projects within the ONS.

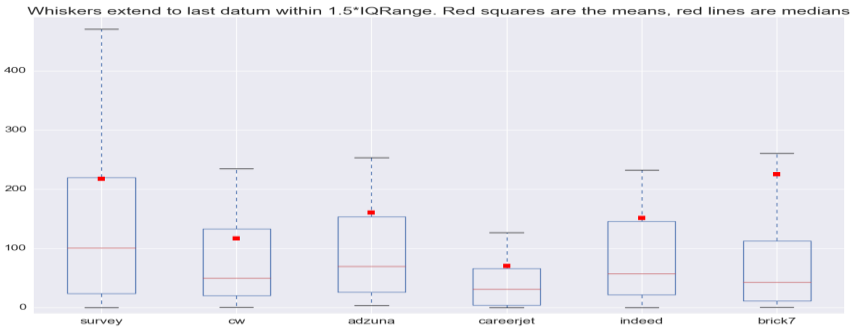
4. Methodology

4.1 Quality : Comparison of data sources

This section focuses on the differences in vacancy counts between various on-line sources for the sample of 50 companies in the JVS captured on 7 April 2017, the reference date for the April 2017 JVS. The combination of a small sample and a high proportion of cases where there are large differences between the JVS and on-line sources (referred to as outliers) means that this analysis should be treated with caution. However, this comparison helps with identifying some key issues with the sources and getting an indication of the scale of some of them. It also gives an indication of how the various on-line sources perform relative to each other. However, the current sample of 50 is too small to produce reliable measures of the coverage of on-line job vacancies.

The data consist of 50 rows, one for each company from the sample. Each row then gives a corresponding JV count from the JVS, the company website (CW) and the 4 job portals (Indeed, Careerjet, Adzuna and Brick7). The analysis focuses mainly on comparing these job portal counts with two benchmarks: i) the JVS (the overall "gold standard"), and ii) CWs (the "online gold standard"). Figure 3 shows the distribution of counts by source follows expectations with the JVS having considerably more job vacancies than any of the online sources. This is because the JVS will include jobs not advertised on-line and because some on-line jobs will be advertised via agencies rather than directly by the employer. The vacancy count distributions for Adzuna, Indeed and Brick7 compare reasonably well with that for CWs, but Careerjet reports lower counts.

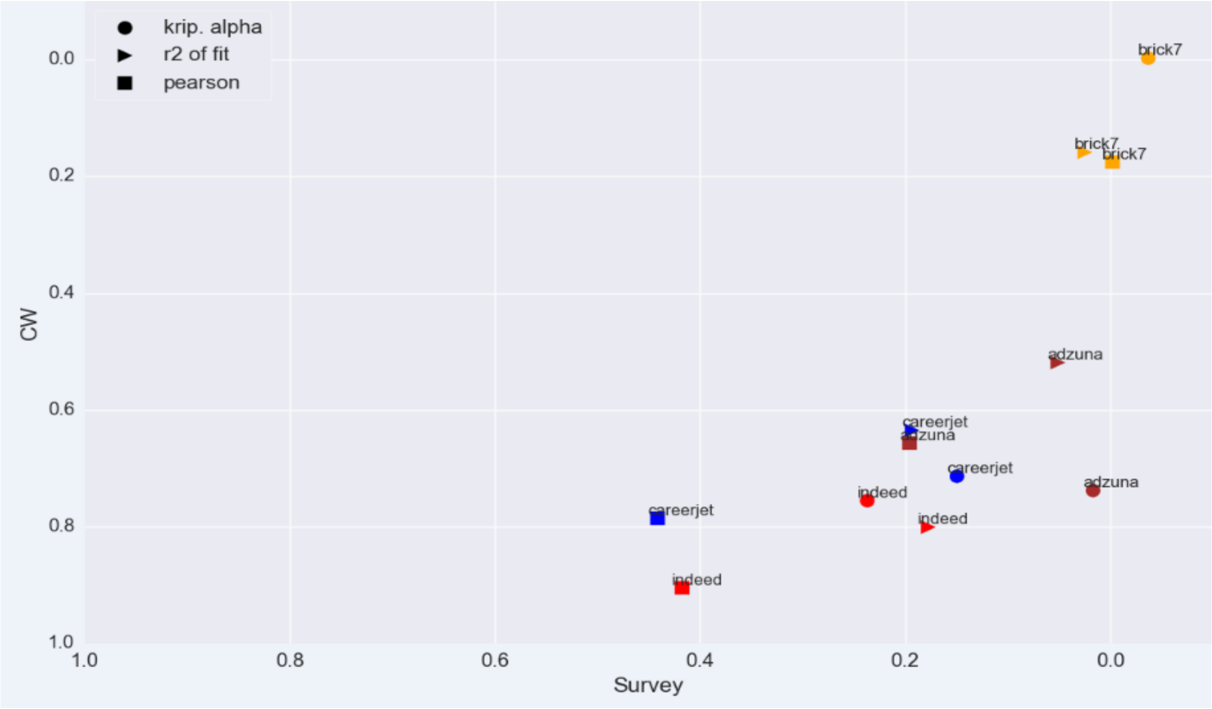
Figure 3: Boxplot of JV counts for each source



We are interested in how well each job portal counts compare with the two "gold standards", the JVS and the CW counts. We started by using the two most common correlation metrics namely, R^2 correlation and Pearson's correlation coefficient. However, these metrics are not strictly correct because this analysis is not investigating explanatory relationships between variables, but rather different measurements of the same variable. Also, a good correlation between two sources may not necessarily produce a good *match*. Krippendorff's alpha is a metric usually used in content analysis and is used to indicate the level of agreement between two measurements. This may be a more "correct" measure of agreement.

We established that a useful way of analyzing how well each portal compared to the two gold standards was to plot the values of each correlation/agreement measure[19] on to a graph with the CW and JVS on the two axes (Figure). All three measures have a maximum value of one (signifying perfect correlation or agreement). Therefore, those job portals with metrics closer to the bottom-left corner are better than those in the top right. Indeed compares best to the company websites followed closely by Careerjet, then Adzuna with Brick7 lagging. Indeed and Careerjet are roughly equal best performance compared with the survey counts, but it is clear that there are much bigger differences generally between the job portals and the survey.

Figure 4: Comparison of job portal counts with counts from company websites and the JVS



The reason for the low correlation/agreement measures between the job portals and the JVS is doe to some very large differences in job vacancy counts between sources for some reporting units, which we refer to these as "outliers". We wanted to investigate what happens when these outliers were removed. It could be argued that outliers should always be included to present a true picture of

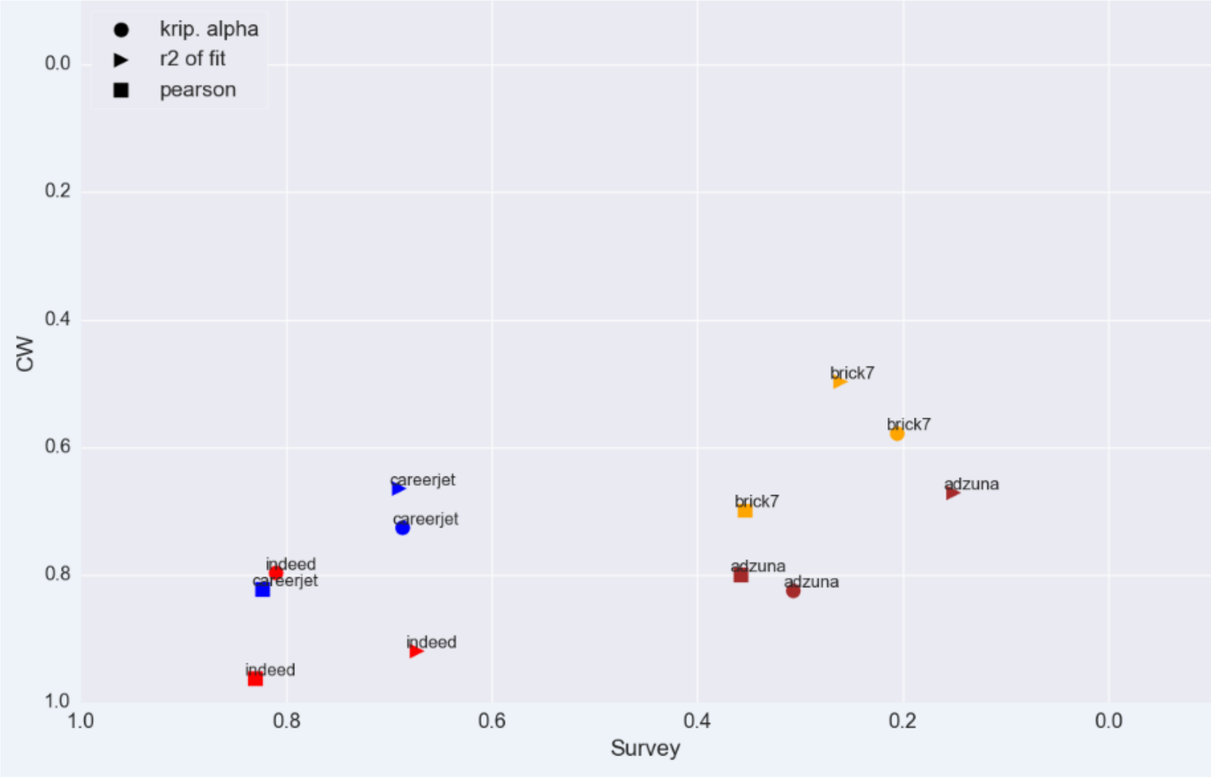
[19] Due to the small sample, values are computed using bootstrapping with 100 iterations

how the two sources correspond. However, one could also argue that these huge differences may exist because the two counts may represent different underlying entities (e.g. a company advertising only its management vacancies online, while reporting all vacancies in the survey). Excluding outliers in such cases may give a better idea of the quality of different sources where the counts should more or less correspond. The following rule was applied:

(x, y > 5) and (x/y > 5 or y/x > 5) ⟹ y is an outlier.

The previous analysis was then repeated with outliers removed (Figure 6). Although the outlier rule is conservative, this made a big improvement in terms of correlation/agreement between the job portals and the survey, especially for Indeed and Careerjet. The removal of outliers did not change the rank order of performance, but, it did seem to reinforce the notion of Indeed and Careerjet performing better against the gold standards than the other two.

Figure 5: Comparison of job portal counts with counts from company websites and the JVS (outliers excluded)



These correlation/agreement measures are a single metric and so do not give a picture of how the distribution of counts compare. Ideally, the distribution would have a zero mean (job vacancy counts are the same on average) and a low standard deviation (differences being usually the same/few outliers). We are also interested, how the differences vary with increasing job vacancy counts.

These questions could be answered by referring to the plots on Figures 6 and 7. These give a visual comparison for the various sources against JVS (outliers removed) and CW (outliers included), respectively. For each pair of sources a Bland-Altman scatter plot is presented showing the difference in corresponding JV counts against their average, with the red line representing the ideal zero difference. A kernel density estimation (KDE) plot then gives an indication of how much the differences cluster around the mean value. As expected CW matches the best with JVS, although Indeed is close behind. A useful feature of a Bland Altman plot is that is clearly shows any systematic bias between different sets of counts. Careerjet looks to be a solid match against CW, but under reports against JVS, especially with higher vacancy counts (i.e. for large companies). Adzuna and Brick7 do not match as well with either of the gold standard benchmarks.

Figure 6: Distribution of count differences between online sources against JVS

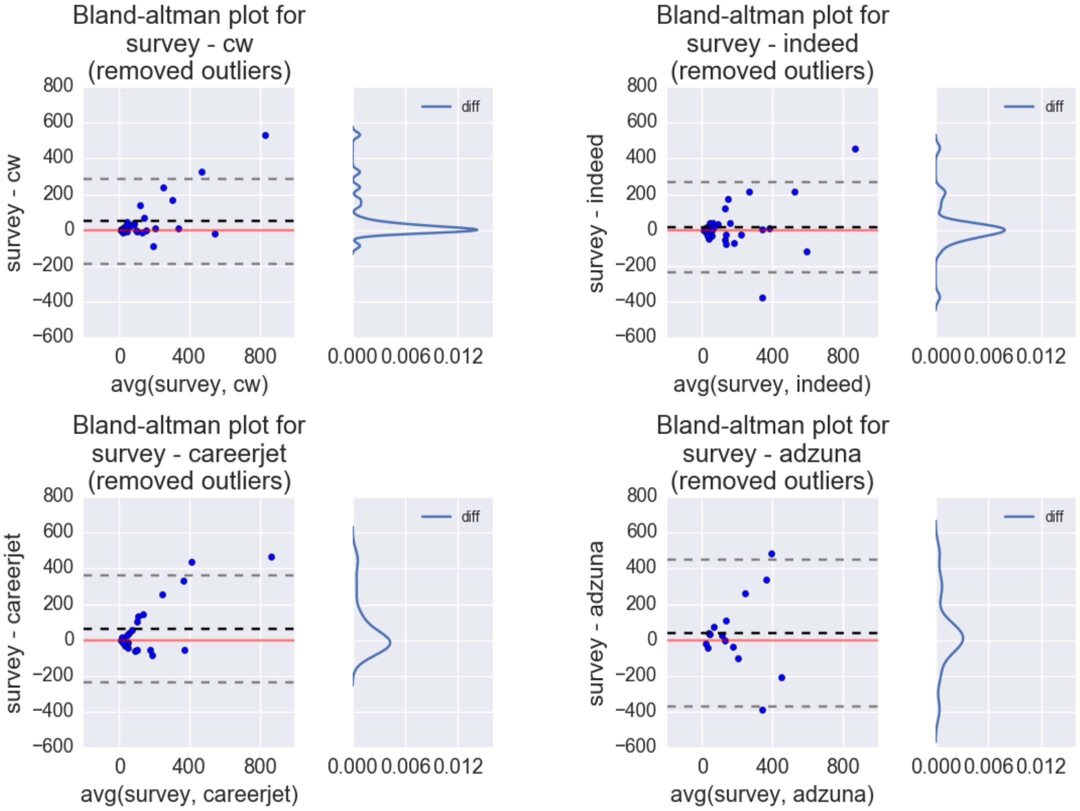
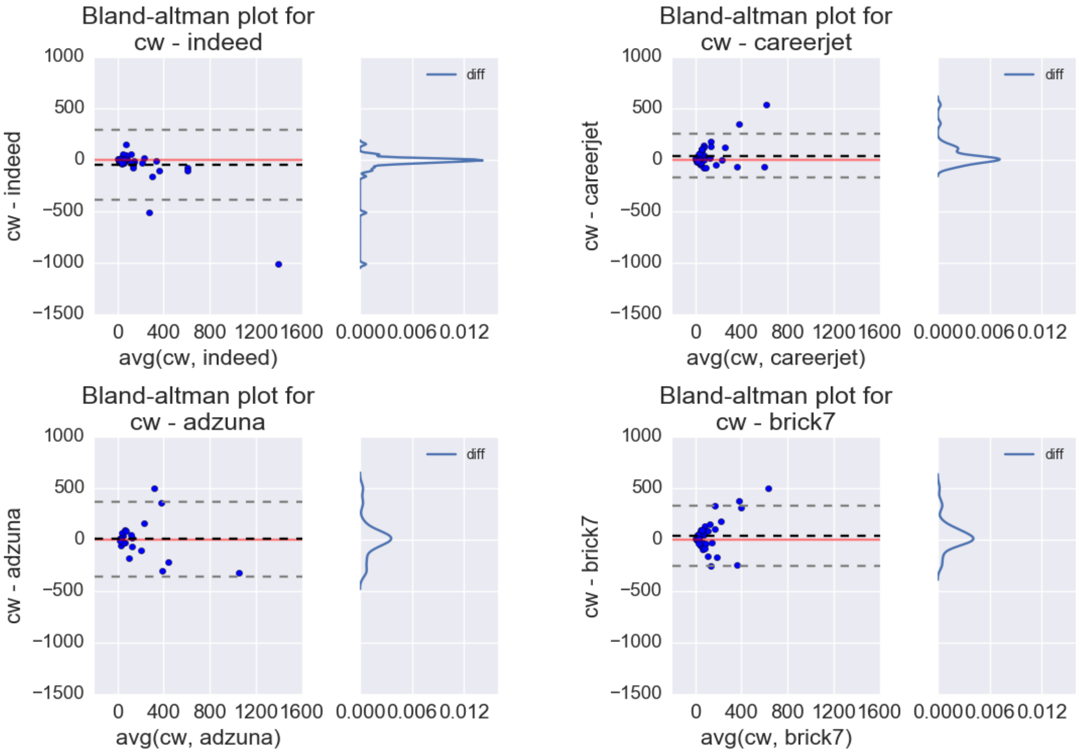
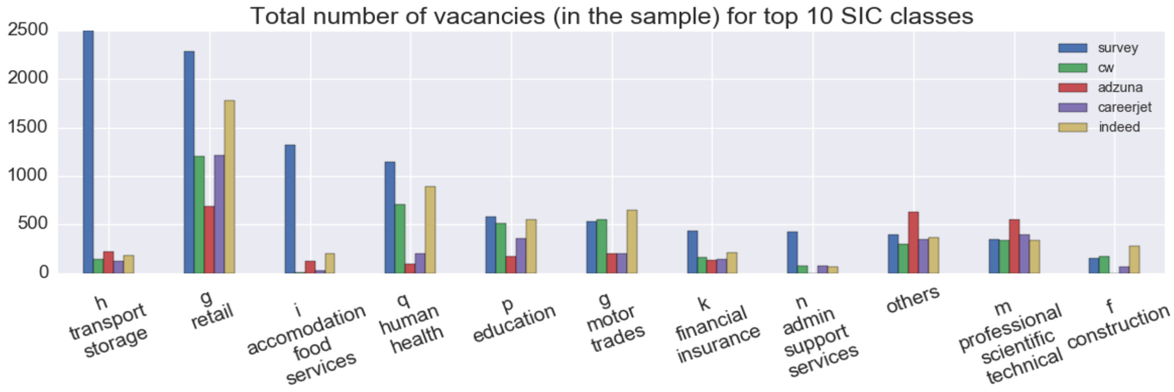


Figure 7: Distribution of count differences from online portals against company websites



A sample of 50 is not really large enough to split the analysis above by the industries of the reporting units. However, a cursory look into the total number of vacancies per SIC class (Figure 8) suggests that SIC classes M (Professional, scientific and technical activities), P (Education), G-98008 (Retail), G-98006 (Motor trades) and Q (Human health) may have the best potential to be captured online. Conversely, groups H (Transport & storage) or I (Accommodation & food services) in our sample had a much fewer vacancies in the on-line sources compared with those in the JVS.

Figure 8: Breakdown of the number of vacancies (from the sample) by SIC class



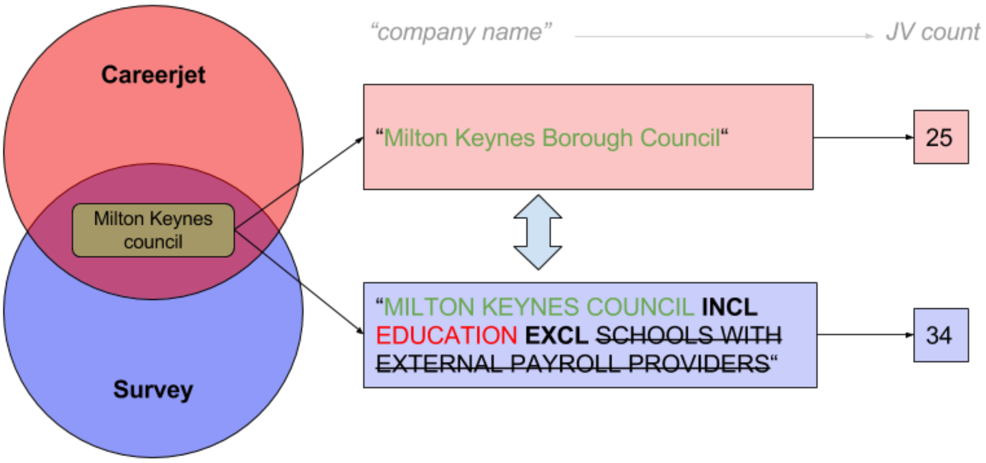
Although this analysis was carried out only on a sample of 50 companies, this can already provide some insight into the quality and completeness of the job vacancy estimates found on the web in general, as well as on individual job portals or company websites. Compared with the ONS survey, the best matching online source was the company websites, thus confirming the assumption that this is the gold standard for on-line sources. Indeed was the best performing job portal against both the JVS and the company websites. Careerjet was the second best performing portal with Adzuna and Brick7 lagging behind. However, no online source by itself achieved a strong correlation or agreement with the JVS unless outliers (typically 25-45% of entries) were removed.

4.2 Matching

In the previous section, we compared different online sources of JV counts based on a sample of 50 companies. There were multiple reasons for choosing only a small sample, including a practical one: the scraped URLs needed to be manually identified for each company and portal, making scaling the comparison up less efficient. An alternative approach would be to take a full set of company counts scraped from a job portal company directory (using the PCP scrapers described in 3.2) and match this against all the reporting units in the JVS. The key challenge here then is to match the company names that occur on the portal to the reporting units from the survey, so that we can compare the counts.

As a case study, we chose Careerjet (CJ), since the company directory on this portal is large enough and easy to scrape, with no restrictions posed in the T&C or robots.txt file. During the period we were running this pilot, 10k-13k company JV counts were found on Careerjet every day, representing 800k-1000k vacancies[20]. Conceptually, there are differences between company names found on a portal and the reporting units (RU) in JVS. There are entries for Careerjet that don't have a matching reporting unit in JVS (e.g. because that RU was not part of that month's sample), and vice versa (company for given RU does not advertise on Careerjet). Our goal is therefore to find the corresponding/matching entries (Careerjet name ↔ RU) for as many companies from the (conceptual) intersection, as possible (Figure 9).

Figure 9: Matching company names


"company name"
JV count
Careerjet
Milton Keynes council
Survey
"Milton Keynes Borough Council"
25
"MILTON KEYNES COUNCIL INCL EDUCATION EXCL ~~SCHOOLS WITH EXTERNAL PAYROLL PROVIDERS~~"
34

[20] Note that this number is actually unrealistically high for such a small set of companies. This is because some of the entries represent other job portals (e.g. an entry for CV-library listing ~150 000 vacancies)

The matching was carried out by assessing text similarity between all possible pairs of <company name on CJ, RU>. The heuristics used included cleaning the entries from stop-words and putting more weight on words with high Inverse document frequency (IDF). However, we should note here that there is a limited chance of success when matching the entries solely on company name, which not always contains enough information. Extra fields, such as address of the company might help make the matching more robust. Unfortunately, this was not available on Careerjet. We thus manually cross-checked the matched entries where there was a higher risk of an incorrect match.

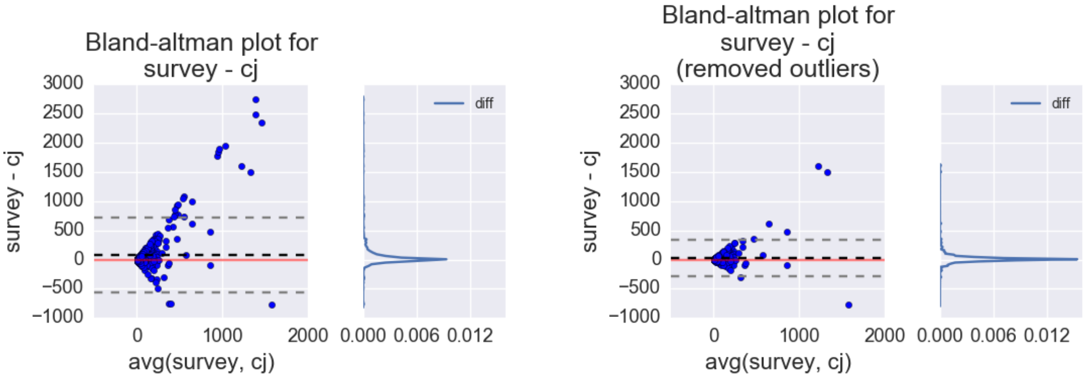
The final dataset consisted of aligned JV counts for 500 reporting units. This number may seem small, considering there were ~11,500 CJ entries and ~11,000 RUs[21]. However, one should keep in mind that the RUs in the survey represent a sample of a much bigger pool of all 2.6 million businesses. And although this sample is stratified by size of the company, most businesses (mainly from lower size bands) are less to advertise on Careerjet. Also, only fairly certain matches were used to avoid false positives.

Careerjet did not compare that well to the JVS in the analysis from previous section. Repeating the comparison with a bigger dataset is confirming this. With outliers included, there was a weak correlation (Pearson 0.32), but even after removing outliers (38% of entries), all the correlation/agreement metrics and had relatively low values (see Table 2 for exact figures). In addition, Bland-Altman plots confirm Careerjet's under-reporting trend for larger companies (Figure 10).

Table 2: Selected metrics comparing Careerjet against JVS

outliers	krippendorff alpha	R^2 of fit	pearson correlation	mean of differences	std. dev of differences	mean of log-ratios	std. dev of log-ratios
included	0.18	0.11	0.32	81.48	316.29	-0.01	3.24
excluded	0.64	0.54	0.73	25.4	142.6	0.32	1.06

Figure 6: Dist (outliers included/removed)



This larger sample of matches provides an opportunity for a better comparison of the breakdown by industry sector (SIC). Table 3 shows the number of entries per SIC along with outlier information. Looking only at those sectors with at least 15 matches, the lowest percentage of outliers was for

[21] Two surveys files were combined to increase the pool of available RUs

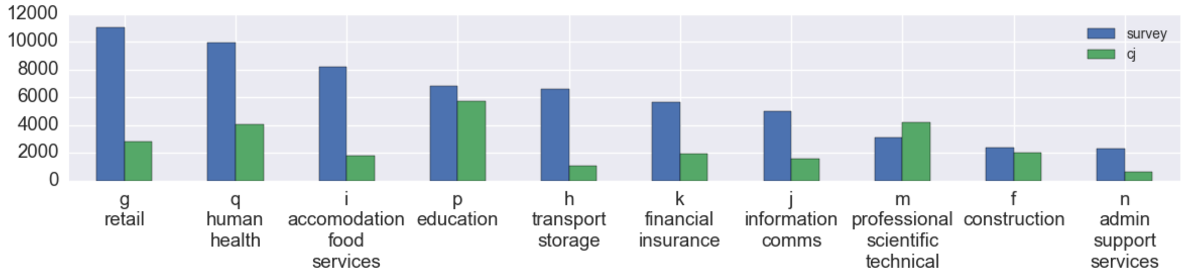
wholesale, education, retail and professional-scientific-technical. The sectors with the highest proportion of outliers were accommodation and food services and administrative support services and transport-storage.

Table 3: Number of entries per SIC sector and corresponding outlier information

SIC Sector	No. of cases	outliers	outliers %
g-wholesale	15	4	26.67
p-education	112	32	28.57
g-retail	44	13	29.55
m-professional-scientific-technical	27	8	29.63
k-financial-insurance	28	9	32.14
o-public-admin	24	9	37.5
c-manufacturing	32	13	40.62
q-human-health	66	29	43.94
j-information-comms	34	18	52.94
h-transport-storage	18	10	55.56
n-admin-support-services	19	11	57.89
i-accommodation-food-services	23	14	60.87

Figure 11 shows the differences in the total number of vacancies per SIC (for the top 10 classes). The two sectors that compared best were professional, scientific and technical services (M) and Education (P). Notable is the difference for SIC class G (Retail), which indicates that although there were only a few outliers, these form a large fraction of the total number of vacancies in the sector.

Figure 7 Total number of vacancies per SIC

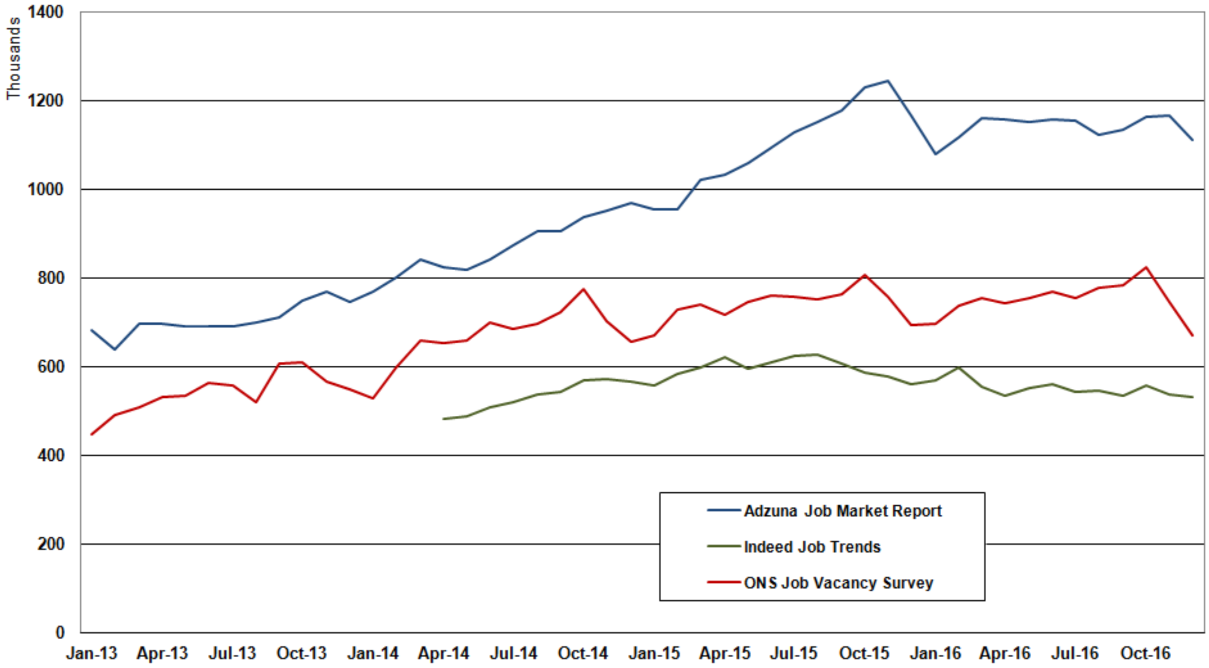


4 Statistical Outputs

This study has made some progress towards two approaches for structuring counts from on-line sources and understanding their quality by benchmarking it against the JVS. However, these frameworks remain very small scale and there are some big differences between various sources that still need to be explored and properly understood. For example, we have not yet looked at job vacancies advertised through via employment agencies. Therefore, it is clear that we are still some way from being able to produce meaningful statistics from these sources.

To further illustrate why producing statistics us difficult, we can compare a time series of total advertised vacancies from two of the largest UK job search engines. Since early 2013, Adzuna have produced a Monthly Job Report containing information about total vacancies, changes in different job categories and associated salary information. From the Indeed website[22], it is has been possible to derive historical data on the total monthly volume of vacancies. These two series are compared with the monthly, non-seasonally adjusted estimates from the ONS JVS (Figure 12). This shows that all three series are quite different both in terms of their levels and trends over time. Adzuna shows the highest level of vacancies and has growth in vacancies during 2015 that is not seen in the other series. Indeed is lower than the ONS but closer than Adzuna. However, neither Adzuna or Indeed seem to have the seasonal patterns that is apparent from the JVS series.

Figure 12: Total job vacancies by selected sources (2013-2016)



This serves to illustrate that statistics cannot be easily produced from on-line sources without understanding their provenance and quality.

[22] This data was taken from the Indeed.co.uk website in late 2016 but

5 Future Perspectives

A key learning point has been that the process of bulk web scraping and cleaning job ads to produce high quality data ready for analysis is very resource intensive and is not a good use of our scarce resources. We are better off investing time in developing partnerships with organisations that already have or plan to acquire these data. This includes continuing our collaboration with CEDEFOP and exploring other partnerships with the support of the ONS commercial data and Procurement teams. We are confident that this lead to acquisition of better quality data within the next few months.

However, the methodological approaches outlined here are still very useful. These have already proved useful in identifying those job portals that have the best coverage. They are helping to understand quality and may also provide the basis for an estimation framework. One aspect we would like to explore further is to see whether by collecting daily counts by reporting unit from a number a portals over a extended period, it would be possible to produce now cast estimates for individual units and then gross these up to produce real-time indicators for the economy as a whole.

This framework may also provide the basis of a service that we could offer to potential partners. By being able to match a partner's data to the JVS and compare counts, we may be able to offer unique insights into their data that no one else can offer. We hope that a by-product of this will be the ability to access other variables and so produce additional analyses on, for examples occupation and geography. Therefore, we see the expansion of the source comparison framework as a priority.