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

Milestone 1.1

Notes from Big Data ESSNet WP1 meeting: 7-8 April (Wiesbaden)

Version 2016-13-07

Prepared by: Nigel Swier (ONS, United Kingdom)

ESSnet co-ordinator:

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

Present:

- Nigel Swier (ONS, Chair),
- Thomas Koerner, Martina Rengers, Susanne Schnorr-Baecker (Destatis)
- Ingegerd Jansson, Oskar Norquist (Statistics Sweden)
- Boro Nikic (SURS)
- Donato Summa (ISTAT)

Apologies:

- Christina Pierrakou (unable to attend due to a strike by Greek air traffic controllers)

1. Introduction:

The main objectives of the meeting were:

1. To share updates on any work in progress

2. To share information on wider developments in the area of web scraping job vacancy data and to consider the impact of these developments on this work package.

3. To develop a strategy for producing concrete estimates, to incorporate this into the work plan and to assign a clear set of tasks for each country.

Since the last face to face (informal) meeting in Luxembourg on 17 Feb, there have been some significant developments. It has become clear that there are a broad range of other actors who have an interest in job vacancy data. These include academic, commercial and other public sector interests. Key players include:

- CEDEFOP (The EU agency for vocational skills training) who have already launched a pilot on web scraping data from job portals.
- Eduworks: Based in the University of Amsterdam and involved in a number of academic research projects using vacancy data from the web to inform public policy.
- Textkernel: A company based in the Netherlands who provide information products based on web scraped data. Some Eduworks research uses data provided through an arrangement with Textkernel.
- There are also other parts of Eurostat and the wider European Commission that have either have an interest in this kind of data (e.g. DG Connect), or have relevant expertise that could be useful (e.g. JRC expertise in text processing)

This has several implications. The first is that there is an opportunity to collaborate with others, share experiences and build on what others have already done (e.g. by reusing tools/programming code and/or possibly by sharing data that has already been collected). The second it is very likely that the level of interest in job vacancy data will in the longer term lead to a more coordinated approach at the European level for obtaining this data and making it available to support public policy.

This means that this WP should not spend too much effort considering the long-term arrangements for gaining access to job vacancy data. Instead effort should be focused on what data is needed to support the objectives of the pilot – namely, how these data could be used within the production of official statistics. The progress already made by others also provides an opportunity to accelerate progress to enabling the WP to put more effort into tackling methodological issues.

2. General Updates:

Nigel gave a brief update on relevant agenda items presented to the ESS Big Data Task Force on 18-19 February, including:

- Research on job vacancy data from CEDEFOP (Vladimir Kvetan) and Eduworks (Pablo de Pedraza)
- The UN principles for use of commercial data. This may be relevant in the context of obtaining data directly from the owners of job portal websites.

Martina gave a presentation on a workshop on mining vacancy data hosted by Eduworks (which included both Vladimir and Pablo as speakers). This focused on the various speakers and their research interests. This could be useful if WP1 identifies a need for certain specific expertise. This workshop was also valuable in terms of understanding the motivations of various stakeholders. In the main, these are focused on improving job matching, that is, filling vacancies more quickly, with the right people. Although these are a slightly different set of interests, with a greater focus on the relationship between job titles and skills, there is very close alignment of interests around the collecting, cleaning and classifying this kind of data to produce data sets for analysis purposes.

The keynote speaker at this workshop was Jakub Zavrel, the CEO of Textkernel. TK have more than 12 years experience in web scraping job vacancy data. They are quite a large operation based in the Netherlands employing over 75 researchers serving over 600 customers. They target a very large number of websites, including it seems thousands of enterprise websites (this could be of interest to WP2). There were also some interesting metrics such as the ratio of job advertisements to vacancies (i.e. duplicates) and how this varies for different countries (in the UK the ratio is 3:1).

Discussion points:

- The Textkernel methods seem to be very complex and it is likely to be difficult to get sufficient transparency. With CEDEFOP it should be easier to understand how the data has been transformed.
- We need to have some idea of the quality. Perhaps we could consider getting a sample of Textkernel data and try to replicate it?
- We should still expend some effort in the collection and processing of data so we understand what is involved.

Action: Nigel to contact Textkernel to request a copy of the Textkernel, request more information on the work they have done on multi-lingual taxonomies and explore the possibility of obtaining some samples of data for benchmarking purposes.

Donato gave a presentation on the updates from WP1 as well as an overview of the web scraping approaches already developed by ISTAT. There are a number of use cases that are being explored by the WP. One of these includes identification of enterprises with job vacancies (although not the number of vacancies or any other information).

Identifying URLs and matching to an enterprise is a key challenge for WP2. One approach being piloted returns a list of possible URLs from the search engine Bing and machine learning to identify the most likely URL for the enterprise. An alternative (or complimentary?) approach could to use a Crowdsourcing platform (e.g. CrowdSearcher) as a way of identifying the correct URL from a list of possibilities.

The "Bing" method is able to identify 100 URLs in about 5 minutes but this depends on many factors. An issue is that official websites may change their URL. One suggestion was to refine the approach by searching within a most relevant subset of URL key words (e.g. "about us" or "contact").

In Slovenia, telecom providers have some information on URLs. In the Netherlands this data is available for purchase. Another approach would be to simply ask survey respondents to decide whether they want to provide a URL or to send a questionnaire. This would both ensure high quality data and would provide legal clarity.

3. Legal Issues

There are legal issues to be considered with respect to web scraping but different countries within the work package have received different legal advice. The legal advice for Germany is that there is no specific law against web scraping and that the terms and conditions of websites around web scraping are not legally enforceable. Nevertheless, the general recommendation is to try to get the consent of the portal owners before using their data on a larger scale. The CEDEFOP pilot also considered the legal aspects of web scraping and concluded that web scraping was more of an ethical issue than a legal one. However, as a public agency it was thought best to seek permission directly from owners of job portals.

Sweden have not yet received clearance to undertake any form of web scraping. There have even been concerns raised around the sharing of job portal data willingly provided by the Swedish Employment Agency (although this was eventually allowed). In addition, an EU Commission report (circulated prior to the meeting) into the use of the internet as a source of data collection provides a legal opinion on the legal feasibility of web scraping. This focuses on the "sui generis" database right, which is set out in Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. This seems to suggest the act of creating a database from web scraped data could breach copyright provisions asserted over that data, at least if essential parts of the database are scraped.

In summary, the legal situation with respect to web scraping remains unclear. However, from an ethical perspective, it would seem to be good practice to obtain explicit permission from job portal owners before doing web scraping. A benefit of this approach is that the job portal owner may agree

to provide more convenient access to the data, and possibly access to historical data. There could be exceptions to this rule, such as small scale web scraping experiments to identify whether scraping a website is indeed possible and to get a small sample to assess the utility of the data before approaching the job portal owners.

Action: Thomas to obtain and circulate a summary of DESTATIS legal advice on web scraping.

4. Technology/Sandbox:

Prior to the meeting, Tony Virgillito (ISTAT) had provided the ESSNet CG with a brief overview of the environment. This provides a virtual environment to enable users to undertake big data experiment developed for the UNECE project. The protocols for members of the ESSNet to access the Sandbox are being developed and will be circulated shortly. There were a couple of questions that Nigel agreed to forward to Peter.

Action: Nigel to ask Peter if there a way for people not directly involved in the ESSNet to use the Sandbox or is it only for the ESSNet? Also are there any arrangements being for training in the use of the Sandbox environment?

5. Interim CEDEFOP Report

Following the ESS Big Data TF meeting in February, CEDEFOP forwarded an interim report on their job vacancy web scraping pilot. The final report is due shortly, but permission was granted to circulate this report to WP members prior to the meeting for discussion on the condition this was not circulated beyond those involved in the WP. This report sets out a detailed methodology for scraping job vacancy websites and preparing these data for analysis. This knowledge is hugely valuable and could save WP1 many months of effort.

CEDEFOP have also indicated a willingness to collaborate with WP1 in other ways. This could include providing code/tools, or even data, since both Germany and the UK were part of the initial pilot. A suggestion from CEDEFOP was to arrange a visit (i.e. to Thessaloniki) to see the live proto-type system and determine what aspects of their pilot would be most useful. It was agreed that we should arrange to talk to CEDEFOP first before deciding on next steps.

There was some discussion about whether there might be some areas where WP1 might be able to collaborate to improve their existing processes (e.g. using APIs where available instead of web scraping, NoSQL database technologies, more application of machine learning approaches). There were some doubts raised about the merits of APIs and NoSQL databases for improving on the existing approaches, but it was agreed that we should continue to explore avenues for collaborating with CEDEFOP.

Action: Nigel to contact CEDEFOP (Vladimir Kvetan) and invite him to a Webex call with WP1 members.

Action: All work package members to think of questions or topics for discussion in advance of the meeting.

6. Country Presentations:

Nigel presented a simple web scraping experiment designed to capture daily vacancy totals by job category from Totaljobs.com, some early exploration of job portal APIs and a comparison of official job vacancy estimates with alternative figures produced by Adzuna in a monthly job report. There is a degree of comparability between these estimates and while differing definitions can plausibly explain some of the differences between the figures, the gap seems to be widening over time. This suggests some possible list inflation with the Adzuna data.

Boro presented some results using the crawler functionality from Import.IO. Slovenia has 31 job portals and 2 main ones. Sometimes the crawler doesn't work or doesn't collect all the required information and so back up approaches may be needed. However, for the purposes of this pilot the approach is promising as a lot of data can be obtained with minimal effort.

Action: Boro to produce some guidance on the use of the web crawlers using Import.io

Action: Nigel to check whether ONS may have already produced some guidance

Action: Nigel to check whether Import.io has been installed on the ONS network.

Ingegerd presented an early assessment of job vacancy data recently obtained from the National Employment Agency. This consists of 4 years of data (over 3 million job ads) and active ads are accessible through an API. The initial results compare favourably with the official job vacancy estimates.

Christina had also prepared a presentation which showed encouraging results using Import.IO using a similar approach to Slovenia.

7. Task Updates

7.1 Template for job portals

Martina presented work on developing a template for evaluating job portals. The initial approach involved reviewing websites that rank job portals. There are about 1600 portals in Germany although this includes regional portals (e.g. for regional newspapers). These should probably be excluded.

A distinction can be made between job portals (which contain original job postings) and job search engines (which crawl job portals) and link search results to the original posting. It was agreed that it would be better to focus efforts on portals, since this would likely reduce issues of duplication.

It was suggested that job portals could be motivated to provide data free of charge by naming them as a source on which official job vacancy statistics were based. This kind of arrangement would have

the advantage of satisfying any legal and ethical requirements and would also likely involve data to be served through more direct means.

There was general agreement on the target variables that should be collected from each portal. The UK has identified a nbew business requirement for information on salary which is often present on UK job ads, although this seems to be less common in other countries.

The consensus was that the aim should be to identify the configuration of portals that would provide the maximum utility overall, while also aiming to minimise the number of target portals. For example, it might be decided to target a few of the best large portals and then supplement these with specialised portals to ensure adequate coverage.

Action: All countries to apply the template define the most important job portals

7.2 Concepts and Practices

Ingegerd provided an update on work to define concepts and comparing practices between countries. Although Eurostat set out minimum requirements for job vacancy statistics, the outputs vary considerably across member states. These may include differences in what variables are collected and the target population (i.e. whether the survey targets legal or local units). The WP pilot will need to be able to derive the correct reference period from any job vacancy data to be able to match to the samples and the evidence from CEDEFOP suggests that this will be straightforward. An earlier UNECE report outlines some these differences from an earlier study, but this framework will require some further elaboration.

Boro made a strong case for using the business register and job vacancy survey samples as the basis for evaluating the coverage of job portals and as the basis for developing a method for producing job vacancy statistics. Depending on the required outputs, it may not be necessary to collect all or even most job ads. The more important thing is have sufficient coverage by industry and job sectors. Understanding the relationships between the data from job portals, the existing survey data and business registers may provide a sufficient basis for producing reliable estimates.

Further discussion points:

- The variables used for the weighting scheme need to be determined (e.g. do we use admin data?)
- We need to better understand how the business areas would like to improve job vacancy statistics.
- Some variables cannot be replaced by data from job portals (e.g. Sweden collect information on when vacancies are filled)
- We will need to explore how well we can match the enterprise name from the job advert to the unit on the business register. It was decided that we should exclude jobs advertised through employment agencies and worry about these later.
- It should be possible to derive reference periods fairly easily. The data model used in the CEDFOP pilot includes opening and closing dates for job vacancies.

Action: All countries to document current practices for each country / identify job vacancy definitions used and compare with Eurostat definition.

Action: Boro/Ingegerd to circulate material from UNECE project; Ingegerd to provide a template for the documentation of the national job vacancy surveys

Action: All countries to obtain business register samples

8. Outline strategy

The following outline strategy was agreed:

1. Apply the process developed by Germany for assessing job portals

2. Review the current practices and conceptual definitions within each country for producing job vacancy statistics (using the common template)

3. Design experiments using job portal data to replicate as closely as possible the current job vacancy outputs produced by each country.

4. Explore the potential for producing new and or improved new outputs using job portal data. This may require collecting additional data.

It was agreed that all countries would seek to broadly follow this approach, although the specific approach may be tailored to the circumstances in each country. For example, Sweden expects to focus primarily on the data already obtained through the National Employment Agency. Germany and the UK may be able to use data collected by CEDEFOP. It was suggested that we should approach job portals to see if we could get historical data.

One useful preparatory exercise would be to perform a "stress test" on some large job portals, to establish whether it is possible to web scrape data without being blocked.

Action: UK/Slovenia to discuss testing large scale job portals.

9. Next meeting

It was suggested that the focus of the second face to face meeting due in Autumn 2016 could be more focused on solving specific technical problems and identifying solutions. This could take the form of a sprint. One option could be to have the meeting at ONS in Newport which has the facilities to host this kind of activity. An alternative approach might be to arrange a "virtual sprint" which would take place over 2-3 days. This could take the form of Webex calls to agree objectives and tasks, breaking off to work on assigned tasks and periodically reconvening to review progress. This kind of approach has already been trialled within the ONS Big Data team with some success. It was agreed that this is something that we could try out at some point over the summer.

The consensus is that now we have a clearer sense of direction, monthly Webex calls to monitor progress would probably be sufficient.

Action: Nigel to organise a virtual sprint at some point over the summer

Action: Nigel to book monthly WP1 Webex calls to monitor progress.