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 and 2

Web scraping / Job vacancies

We scraping / Enterprise Websites

Minutes of joint meeting in Rome 7-9 November 2016

Version 2016-23-11

Prepared by:

Moncia Scannapieco (ISTAT)

Nigel Swier (ONS)

ESSnet co-ordinator:

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

Day 1 (7/11/2016): WP1 meeting

Participants

Nigel Swier –ONS
Thomas Korner -DESTATIS
Martina Rengers DESTATIS
Boro Nikic SURS
Cristina Pierrakou - ELSTAT
Ingegerd Jansson - SCB
Dan Wu - SCB
Olav ten Bosch- CBS
Monica Scannapieco – Istat
Donato Summa- Istat
Stefano De Francisci – Istat
Alessandra Righi- Istat

The planned agenda for this meeting day was:

1. Welcome / Introduction / Overview	10:30
2. Country updates (format 15-20 minutes presentation followed by 10 minutes discussion)	10:45
Lunch	13:00
3. Country updates (format 15-20 minutes presentation followed by 10 minutes discussion)	14:00
4. Review WP1 Deliverable 1.2 (Interim Technical Report)	14:45
5. Planning the rest of SGA-1	16:15
6. SGA-2 planning	16:45
7. Review of future events / conferences	17:15
Close	17:30

1. Welcome / Introduction

- Nigel welcomed participants to the meeting. This meeting is just after the half-way point of SGA-1 and a third of the way through the ESSNet as a whole so this is a good opportunity to review and assess progress.

2/3. Country Updates

United Kingdom:

- ONS experiments with de-duplicating job ads using supervised machine learning indicate that this may be effective but would require a large amount of clerical resource to implement properly.
- Coverage assessment is being focused on matching job ads the 1300 large businesses that are always in the JV survey. This is proving difficult due complex ownership structures and differences between the name of reporting units and company names and so auxiliary data from the BR may be needed.
- There have been issues in working with the data provided by CEDEFOP. CEDEFOP are starting with the next phase so this may be an opportunity to review and strengthen our cooperation.
- Several ONS staff working on this pilot have left over the past few months which is effecting progress.

Germany:

- Job portal landscape in Germany is very complex, with a large number of portals that are constantly changing. Analysis of 'hybrid' portals shows large differences in the percentage of original job ads as a proportion of all ads and also differences in the number of duplicate ads. The Federal Employment Agency (FEA) has by far the largest number of original jobs and so may be the single best source.
- A comparison of aggregate job ads (from Stepstone) by industry against equivalent data from the JV survey shows that the industry profiles are very different and so on-line job ads are clearly not representative of all job vacancies.
- The 2015 JV survey asked some supplementary questions about recruiting channels, which will be very useful in understanding difference between employers, particularly in terms of enterprise size
- The job portal of the FEA (Jobborse) is the largest. A meeting has been arranged with the FEA to better understand the differences between this source, a separate register based statistic, and the JV survey.

Greece:

- Greece has experimented with web scraping two job portals. Two tools have been used: i) Import.IO ii) Content Grabber. The latter has some advantages (e.g. error handling). It is possible to scrape data from each portal without duplication (but further work is needed to remove duplicates from the joint data set.
- A matching exercise has been made to both the job vacancy survey and the business register. From the sample of 3060 ads, the employing enterprise could be identified in 55% of cases and of these, the job ads were advertised by 256 different enterprises. 9% of these enterprises could be matched to the JV survey and 30% to the Business Register. Of the 45% of job ads where the enterprise name is not identified, it is believed that many of these are job vacancies.

Slovenia

- Good progress has been made in developing a prototype system for scraping job ads from the two largest job portals deduplicating them and the producing a weekly time series.
- A multi-dimensional approach has been developed for matching company names form job ads to the business register
- Coverage assessment of the scraped job portal data against the JV survey shows that about 25% of vacancies are advertised on job portals. However, if ads from enterprise websites and public sector jobs (derived from admin data) are included, then coverage is 54%.
- Future activities include:
    - Additional questions on 2017 JV survey on advertising channels
    - Meetings with Slovenian Employment Agency and Private Employment Agencies
    - Analysis by economic activity and enterprise size, development and testing of estimation models, and use of occupation.

Sweden

- Work has been focused on linking Swedish Employment Agency (PB) data, JV survey and the business register.
- The quality of the PB data is generally pretty good. It records the organisation id so it can be linked to business register. However, there are a lot of outsourcing and recruitment agencies so this may not always be accurate.
- For matching on local units it has proved to be easier for public sector (71%) rather than private sector (14%). It may be possible to improve matching using address information.

4. Review of interim technical report:

- There was general agreement about the scope and positioning of the report and the need to emphasise to the review board that this was a work in progress.
- Some country sections for the interim report already received.
- For the final SGA-1 report we should update the section on JV survey to follow template developed by Sweden.

Action: Outstanding contributions should be sent to Nigel by the end of the week (i.e. 11/12/16)

5. Planning for rest of SGA-1:

- It was agreed that the virtual sprints were a good idea and that we should aim to have a 3rd sprint around the end of Jan beginning of February. It was also proposed that we experiment with using the Webex facilities for teleconferencing (i.e. without video) to see if this resolves issues with sound quality. Possible topics include another sprint on matching or development of a quality framework.
- There was discussion about whether web scraping was the best long-term approach for accessing data and that it might be better to explore gaining direct access to data from job portal owners. There was also agreement to look at reviewing our links with CEDEFOP and looking at how we could support their pilot. It was proposed that a possible future sprint could involve brainstorming on partnerships and what we could offer to job portal owners to give us access to data. It was

agreed to explore the possibility of doing something face to face before the Sofia dissemination meeting as all participating countries would be present.

Action: Nigel to organise a meeting in Sofia to discuss partnerships

- It was noted that the time scales for producing the final technical report meant that we needed at outline structure fairly soon,

Action: Nigel to propose an outline structure of the final SGA-1 report and to circulate.

6. SGA-2 planning

- A key issue for SGA-2 is how to incorporate the four new partners. It was agreed that the approach should be to set up some kind of framework or set of guidelines that these countries can follow so the aim would be to test these guidelines. It was suggested that one way of starting to produce some guidelines would be for each country to propose 10 elements of guidance from which we could compile a single set of draft guidelines. This will need to be produced as part of the final SGA-1 technical report.

Action: Each country to propose 10 elements of guidance

- An important element for SGA-2 will be exploring the collection of information about job vacancies collected for enterprise websites using the tools developed by WP2. This will be covered at the joint meeting tomorrow
- An early activity in SGA-2 will be a workshop which will review the work from SGA-1 and will help to bring the new partners into WP1. There are three possible venues: Newport/Cardiff, Ljubljana, and Wiesbaden.

Action: Nigel, Boro and Thomas to get more details on the feasibility of hosting a workshop in September 2017

7. Review of future events/conferences

- There are several events coming up the involve presenting work from this WP (i.e. Towards Agile Social Statistics, November 2016, ISI July 2017). It would be good practice to make sure all participants are aware of such events and to share any presentations beforehand.

Action: Nigel to share presentation and paper for the Towards Agile Statistics Conference.

Day 2 (8/11/2016): Joint WP1/WP2 meeting

Participants

Nigel Swier –ONS
Thomas Korner -DESTATIS
Martina Rengers DESTATIS
Boro Nikic SURS
Cristina Pierrakou - ELSTAT
Ingegerd Jansson - SCB
Dan Wu - SCB
Olav ten Bosch- CBS
Kostadin Georgiev -BNSI
Galia Stateva, BNSI
Jacek Maślankowski- GUS
Matthew Greenaway – ONS
Monica Scannapieco – Istat
Donato Summa- Istat
Stefano De Francisci – Istat
Giulio Barcaroli- Istat
Alessandra Righi- Istat

The planned agenda for this meeting day was:

1. Welcome / Introduction / Overview	10:30
2. Legal Issues / Review of WP2 Legal Report –all	10:45
3. Discussion on how to go on with the Legal Report and Wp1 contribution to it- All	11:20
4. Demonstration of WP2 system -Istat	12:00
Lunch	13:00
5. Demonstration of WP2 URLs retrieval-BNSI	14:00
6 Demonstration of WP2 related software –NL	14:45
7. Demonstration of WP2 related software –PL	15:30
8. Review of WP1/WP2 dependencies- All	16:15
Close of joint meeting	17:00

This agenda was re-scheduled given that some participants had to leave in the afternoon. In particular, the demonstration session (points 4,5,6,7) was anticipated to the morning.

Demonstration of a system for URLs retrieval and Web scraping – Istat

Giulio Barcaroli and Donato Summa had a joint presentation on the Istat software system for URLs retrieval and Web scraping.

The presented system includes both ad-hoc developed software as well as a proposal of a methodologic approach to accessing and processing Web sites contents.

Comment: there was a discussion on the possible language specific features of the system, concluding that the system is configurable for different languages.

Demonstration of a system for URLs retrieval-BNSI

Kostadin Georgiev and Galia Stateva presented a URLs retrieval solution developed at BNSI.

The main features of such a solution include:

- Use of two search engines (JABSE and GOOGLE)
- Dealing with multiple character sets (Bulgarian and English)
- Interface for manual inspection and comparison of the results from the two searches.

Galia also showed a video about the results of the running of Istat's URLs searcher on BNSI enterprises.

The URLs searcher was run but the results, both in terms of content and visualization, were hard to check due to problems with the Cyrillic alphabet.

Action: Istat to follow up with BNSI for solving the problem.

Presentation of WP2 related software systems – CBS

Olav ten Bosch presented a URLs retrieval solution that was experimented at CBS.

Main features of the solution:

- Sample of 1000 from enterprises with URLs
- Use of Google APIs-- Search4Stats: Generic architecture available on GitHub
- Blacklisting
- Use of Snippets
- Python scripts for the learning task.

Comment: need to take into account also Enterprise with no URLs

Presentation of WP2 related software systems – GUS

Jacek Maślankowski showed a solution for the implementation of the social media use case.

The solution looks for Facebook and Twitter references on the main pages of the enterprises sites. If on the main pages no link is present it goes deeper in the sites. The solution is based on an ad-hoc analysis of the sites' source code to identify links. It is based on Python 3.o and HTML parser.

Comment: useful analysis on comments, e.g. social media as a mean for job advertisement

Review of the Legal report and discussion of WP1 contribution

Each country presented the contribution to the legal report

Actions:

- Need to add a section on EU and International Regulation (Dan)
- Need to share a common Netiquette protocol (Matthew and Olav)
- Coordination with SOGETI consortium (Alessandra and Boro)
- Final section reporting specific legal issues related to WP1 work (Nigel)

Day 3 (9/11/2016): WP2 meeting

Participants

Nigel Swier –ONS
Dan Wu - SCB
Olav ten Bosch- CBS
Kostadin Georgiev -BNSI
Galia Stateva – BNSI
Jacek Maślankowski- GUS
Matthew Greenaway – ONS
Monica Scannapieco – Istat
Donato Summa- Istat
Stefano De Francisci – Istat
Giulio Barcaroli- Istat
Alessandra Righi- Istat

The agenda for this meeting day was:

1. Welcome / Introduction / Overview	9:30
2. State of the work SGA1 and planning of next steps	9:45
3. Structure of the final SGA1 report	11:30
Lunch	12:30
4. SGA2: plan of the work and of the events	14:00
5. AOB	15:00
Close of Wp2 meeting	15.30

After the welcome, Monica Scannapieco presented the item 2 of the agenda, namely the "State of the work in SGA1 and the planning of the next steps".

There was a discussion on the planning of the next steps until February 2017, where there is the milestone of the Sofia meeting.

The principal **decisions** are summarized in Table 1 and were:

- Work on the prototypes for 4 out of the 6 identified use cases
- Each country committed on specific use cases (see Table 1)
- Given to the legal constraints on scraping, some specific decisions regarded Sweden and UK.
    - Sweden will work only on the Job Vacancy use case, especially as a contact point with WP1.
    - UK will work on selected use cases especially on the analysis phase.
- With respect to the two use cases not selected in the work perimeter of SGA1, there is however interest to start sharing some ideas especially from UK and IT.

Table 1: Use cases participation

Use Cases\Countries	IT	SE	UK	NL	BG	PL
1: URLs retrieval	**X**		x	x	x	x
2: Ecommerce	**X**		x	x	x	
3: Job advertisement	x	**X**	x			x
4: Social Media	x				x	**X**

- In order to carry out the work more efficiently, Monica suggested to have responsibilities to coordinate use case prototypes. These were selected on a volunteering base and are shown in Table 1 with a bold cross. In particular they are:
    - Italy for Use cases URls retrieval and Ecommerce
    - Sweden for Use case Job advertisement
    - Poland for Use Case Social Media
- It was decided that the responsible of a Use case will coordinate the work of the pilots related to that use by specific actions (emails or virtual meetings). At the general WP2 meeting the responsible of a Use case will report on the work done on that Use case.

On the agenda item, "Structure of the final SGA1 report", Monica Scannapieco illustrated the following sections proposal:

- General Motivations for Web Scraping of Enterprises Web Sites (WSE-WS)
    - Use cases
- Description of a framework for (WSE-WE)
    - Logical building blocks
- Methodological Issues
    - Specific vs generic scraping: when/what
    - Analysis techniques: machine learning?
    - Comparative analysis of methods used in pilots
- Technological issues
    - Comparative analysis of tech environments used in pilots
    - Issues
- Appendixes by use cases

The proposal was accepted. In particular, it was remarked the importance of having Appendixes "by use case" to be written during the implementation of the use cases pilot.

On the agenda item "SGA2 and planning of the next events", there was first a discussion on SGA2.

Monica Scannapieco shared the revisions made to the text of WP2 for SGA2 according to Eurostat first request for clarifications. They were basically (i) the introduction of a use case to evaluate the possible

usage of the methods and tools developed within the WP2 work for the European Group Register and (ii) the evaluation of applicability of the WP's outputs, in particular considering production scenarios making use of the proposed approaches. There was a commitment by the participants to this new text.

With respect to the next events, both within SGA1 and SGA2, it was decided:

- The two next WP2 general virtual meetings were scheduled for December 12 and January 12.
- There will be a proposal of a back-to back meeting of WP2 in Sofia, in conjunction with the general dissemination meeting.
- In SGA2, there are two physical meetings planned: the first one in October 2017 and the second one in March 2018. There was a general agreement on trying to have the second meeting jointly with WP1.