Minutes
ESSnet Big Data II

To WPC partners ESSnet Big Data II

cc

from Vera Ivanova

subject: **Draft minutes 1st WPC meeting ESSnet BD II by WebEx 23rd of November2018**

2018-11-26

Participants

Galya Stateva (PL WPC) - BG	✓	Olav ten Bosch - NL	✓
Lukas Mikesa - AT	✓	Jacek Maślankowski - PL	✓
Alexander Kowarik - AT	✓	Michał Bis - PL	✓
Johannes Gussenbauer - AT	✓	Martin Wood - UK	✓
Caterina Viviano - IT	✓	Jussi Ritola - FI	✓
Monica Consalvi - IT	✓		
Vera Ivanova – technical assistant - BG	✓		

Agenda

1. Introduction

Galya made a short presentation of herself and the team of BNSI. Then Galya welcomed everybody to present themselves.

The agenda was approved and she proceeded on it. In the current project there are 3 new partner countries – colleagues from AT, DE and FI. DE and FI actually won't participate in the websraping activities and in task 3 but they will contribute in the process of developing the methodological frame in task 2. The current project is a natural continuation of the ESSnet on BD I. The previous ESSnet project achieved at least two main results:

- A methodology, process and software implementations for detecting websites of enterprises (URLs retrieval) based on search engines and machine learning techniques and Methodologies, processes and software implementations for detecting characteristics of enterprises such as E-commerce activities, Social media presence, Job advertisements, NACE code and Sustainability reporting on enterprises' websites.etc;

Some output indicators will be produced and published as experimental statistics:

- Rate of retrieved URLs from an enterprise list
- Rate of enterprises engaged in web sales on their website

- Rate of enterprises with job advertisements on their website
- Rate of enterprises that are present on social media
- Percentage of enterprises using Twitter for a specific purpose, estimated from web data
- These results can already be deployed and implemented in any ESS country but it may require adaptation to the local circumstances to make them effective.

2. Overview of WPC

Galya presented the achieved results from ESSnet on BD I – they are available on project wiki page (https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP2_Webscraping_enterprise_characteristics1).
ESSNet on BD II: the current state is that it is waiting for official approval by Eurostat but it actually started on 1st November 2018.

3. Organization of work and distribution of tasks

In her presentation Galya made a proposal for distribution of tasks and timetable. On asking for some comments or proposals UK and AT agreed with the proposal.
Next in the presentation the deliverables and deadlines for each task were commented.
For Task 1, ESS web-scraping policies: BG will expect the assistance of UK because UK already has their national webscraping policy. Martin agreed and confirmed that it is proper for all webscraping policies and it could be suitable starting point for webscraping policy. He also approved the proposed timetable.
For Task 2, Methodological Framework/Guidelines: Olav asked to send his comments later in an email. AT thinks that 1st and 2nd deliverable (Identification of statistical production processes in terms of four phases of the BD life cycle and possible statistical outputs at national level; Update of the WP2 use-cases and selection of potential new use cases) are too ambitious to be completed in January 2019 and recommended pushing the deadline to February 2019. IT also mentioned end of February 2019 as deadline. All other participants agreed. The Italian colleagues were of the opinion that for other tasks it would be good all countries to participate but for this task it would be better each country to focus on some activities. For task 2 IT would take 3rd and 5th subtasks (Development and test of functional prototypes for collecting, processing and analyzing of webscraping data in order to produce new statistics or enhancing existing ones; Data and application architecture for BD production). Jacek agreed with this conducting the tasks – end of February 2019, so that to have enough time to September 2019 to have functional prototypes tested. Galya proposed activities 3 and 4 (Development and test of functional prototypes for collecting, processing and analyzing

of webscraping data in order to produce new statistics or enhancing existing ones; Definition of the implementation requirements of prototypes in the relevant statistical production processes at national and European level) to be done simultaneously, to change the deadline of activity 3 to September 2019. AT agreed to Galya's proposal. FI had no comments, it looked fine for them.

For Task 3, Experimental Statistics: all participants are included except FI and DE. Galya put the question for the recommended period for web-scraping activity (ICT usage in enterprises, e.g. March-April 2019) because this period is good for comparison after that. Alex (AT) added that data have to be submitted in July 2019. Jacek answered that most of the countries had tested their own solutions, so to ask only AT and UK. Olav asked if it was the situation to repeat more widely old cases for all countries. Galya explained that we had to update/extend the previous WP2 use-cases and not only repeat our work from previous project and to define possible "new" use-cases, e.g. webscraping for small enterprises.

As concerns Task 4 Starter Kit for NSIs, we need to select software for each case. Jacek confirmed the necessity to prepare software for each use-case by 2020. IT answered that for the moment they accept the timetable and would talk with their IT experts.

Regarding Task 5, Quality template for statistical outputs: there are 3 subtasks and for the 1st subtask (Investigation of the UNECE Framework for the Quality of Big data and SGA-2 WP8) Galya asked if the participants agree. AT agreed in principle, but added that quality issues are in scope of WPK, so maybe it would be better to wait for some inputs from the WPK work. Galya also pointed out that it would be better to keep the date as now and after the kick-off meeting in Vienna to change the date according to WPK date.

Galya finished her presentation with illustration of the pipeline of WPC which is almost the same as logical architecture of business process of webscraping, defined under WP 2, ESSnet on BD I. Jacek confirmed that nothing should be changed and these first steps were ok.

Galya pointed out the collaboration of WPC with WPB, WPJ and WPF and asked for ideas for collaboration with WPB and WPJ regarding webscraping techniques – whether and how? According to AT the architecture and software should be the same and the interaction with WPF should be passive. Galya added that after the kick-off meeting in Vienna it would become clearer for the collaboration with WPF.

4. Topics for discussion

Galya asked all the participants after the meeting to send an email with ideas for:

- Old and "new" use cases?
- Functional prototype by use-cases OR…?
- How to achieve economies of scale – e.g. sharing of resources at ESS-level….and is it possible since Eurostat asked us all especially to include a sentence for economy of scale in the updated proposal

WPC webex meeting and F2F meeting – Galya reminded that she had sent a schedule for webex meetings and clarified that she proposed a webex meeting once a month – the Tuesday before the day of CG meeting (according to the CG meeting schedule, provided by Peter Stuijs). Regarding F2F meetings 2 meetings are envisaged – June 2019 and June 2020 – the location will become clear maybe after the kick-off meeting in Vienna, maybe back-to-back with WPB or WPF but we will wait until the kick-off meeting.

Collaboration with Wiki – it is useful for sharing documents and with GiTHub – for sharing software. Martin proposed one repository only for WPC, e.g. WPC wiki.

The next WPC webex meeting is planned for 8th of January 2019.