REPUBLIC OF BULGARIA
NATIONAL STATISTICAL INSTITUTE

www.nsi.bg

**Minutes**
**ESSnet Big Data II**

To WPC partners ESSnet Big Data II
Cc
from Vera Ivanova

subject: **Draft minutes 3rd WPC meeting ESSnet BD II by WebEx 12th of February 2019**
2019-02-18

Participants

| | | | |
|---|---|---|---|
| Galya Stateva (PL WPC) - BG | ✓ | Olav ten Bosch - NL | |
| Alexander Kowarik - AT | ✓ | Jacek Maślankowski - PL | ✓ |
| Martin Wood - UK | ✓ | Jussi Ritola - FI | ✓ |
| Aidan Condron - IE | ✓ | Vera Ivanova – technical assistant - BG | ✓ |
| Caterina Viviano - IT | | Markus Zwick - DE | |

Replacement

| | |
|---|---|
| Eelco van Vilet – NL – replaces Olav ten Bosch | ✓ |

The representatives of IT and DE did not participate in the meeting. Marcus had other professional engagements. ISTAT have organizational changes and Monica had informed Galya that at that moment nobody was responsible for WPC and in 1-2 weeks a new person would be appointed. Eelco van Vilet left the meeting at the middle because he didn't have video on his phone and couldn't be effective in the discussion. Jussi left the meeting earlier due to other professional engagements. Martin joined in the meeting later.

**The proposed agenda is:**

1. The description of the WPC use-cases and statistical indicators: current status

2. ESS webscraping policy

3. Draft structure of the Methodological framework

4. AOB

**1. The description of the WPC use-cases and statistical indicators: current status**

Galya thanked Jacek and Alex for the description of use case 2 and the respective indicators and proceeded on the agenda. She reminded the participants of the conclusion of the previous meeting to have draft descriptions of the use cases by the end of February. She announced that Olav would give a feedback for use case 1 later. Now the very first draft of use case 1 is at hand, elaborated by BG and of use case 2 elaborated by Alex and Jacek.

Galya repeated there are 4 use cases but only the first two cases were discussed because Martin and IT were not here. Galya presented briefly use case 1. She explained that she tried to define the reference population – this is the target/main population for ICT survey according to the Regulation. It should be taken into account that here are included all enterprises, not only enterprises with more than 10 employees, but also small enterprise with 0-9 employees.

In the discussion the opinions were the following:

According to Jacek now the population differs from the population of ICT survey. It's difficult to identify small enterprises because they rarely have websites. He advised to take this step, but for the indicators it would be better to exclude enterprises with less than 10 employees.

Galya agreed with Jacek that small enterprises are excluded from the scope of ICT survey. It would be interesting to see this case.

Alex proposed to divide use case 1 into 2 parts: first part – more than 10 employees and second part – less than 10 employees.

Aidan's opinion is that the case with enterprises with below 10 employees is more experimental, in this case there is more noise in the data. He explained that single person's businesses dominate, like e.g.taxis, and they don't have websites. Statistically speaking businesses are not enterprises.

Alex stated that 2 persons is an enterprise. He suggested excluding of small enterprises from NACE classification. Lawyers are businesses. Some enterprises can be done, others not, it depends on the branch, it's specific.

Galya will update the description of use case 1 after the meeting in accordance to the discussion.

Galya presented the use case 1 indicatorsthey are more detailed list of indicators, of which the last one for NACE categories is an optional indicator. Galya put the question if it would be useful.

Jacek approved having more indicators and suggested the optional one be mandatory. Regarding other indicators his proposal is it to be one indicator with several attributes or dimensions, e. g. to split the indicator according to the number of employees.

Alex agreed with Jacek for more detailed dimensions of the indicator for enterprises having websites.

Galya suggested that for small enterprises it be optional and for big – obligatory. She asked the participants about their opinions about splitting – to split the indicator or to stay in one indicator with different dimension.

Alex stated it would be better in one column all dimensions to be put. Galya reminded that it was a first draft proposal and time is needed to clarify it with Olav and other Dutch colleagues. At the end of February or beginning of March the final version of description of use case 1 would be provided to the WPC partners.

Jacek presented the joint work with Alex on **use case 2**. He stated that population would be defined later. In this use case there are only general indicators and it would be considered if there's need to be specific about the number of employees. According to Galya it's a good idea to include the proposed new indicator – Companies working on upcoming/new phenomena on the example of AI and ML. Galya asked about the variable User behaviour discussed in Vienna and proposed by Jacek, but not included here. Alex said it was forgotten here and Jacek clarified that it depends on cookies, which the user can see on their computer. He also added that indicator 4g Advertisement of open job positions or online job application was taken from ICT survey.

Galya put the question which approach to choose about indicators since she was in favour of more indicators, while Alex and Jacek preferred less because the risk should be minimized with smaller number of indicators. Aidan stated it would be better to promise less and deliver more. Galya noticed that Eurostat would be happier with more indicators but accepted the general preference for less indicators with more different dimensions.

Meanwhile Martin joined and presented shortly the **use case 4**. He explained it was simpler and he had listed primary indicators generated from natural language methods. The text description is what clusters represent. He will send the document later by e-mail. According to Aidan it's experimental and Alex is of the opinion that it's a good experiment and some input to be put in that is reasonable.

Galya asked the partners if it was possible to have more detailed description of use-cases and list of indicators by end of February and communicate them between all WPC countries to discuss it at the next web-ex meeting. Everybody agreed with the deadline.

Jacek requested a discussion on the other sheet – the use case 2 template and section "Actors", especially about the necessity to have such technical details in the list of actors at that moment. Galya confirmed the technical details should stay here in the description of the use cases. Alex explained that in Statistics Austria they would do it with other IT tools. Jacek proposed regarding Apache Spark not to include the specific platform. Galya suggested it would be better to mention the platform, but more general, such as technical platform used in Big data ecosystem. Aidan proposed to use the Sandbox for this aim and Jacek and Martin agreed there is no need to specify the platform and the technology.

Jacek asked about the preparedness of the partners for the massive webscraping starting in March-April with view to the indicators. Everybody confirmed their readiness.

Galya asked Alex if they had chosen the software. Alex said that acquisition of enterprises URLs via Yellow-Pages data base and general web search tool (e.g. Bing or Google) will be done. AT had tested IT software but there were some blocking or slowing down of all processes and they are waiting for Dutch software.

**2. ESS webscraping policy**

Galya thanked Martin and presented in short his work: he made a draft of the webscraping policy. He created a page for it at wiki page, WPC section https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Web_Scraping_Policy_Draft. The following week BG expects a feedback from its legal department so BG will contribute to the preparation of the Policy. So far there is a little bit of an introduction, some notes on how it might work with different NSI's legal requirements, and some section headings to define a structure to be filled. The Netiquette document should provide a lot of the meaningful content.

Galya announces Finland was interested to contribute in the developing process of the Policy. Galya had provided Jussi with all related information regarding webscraping policy and at the later stage the Finnish colleagues intend to contribute on it. Martin explained he had tried to keep it general so everyone who wants to use it has to consider its own policy. It is in full compliance with the European Statistics Code of Practice and GDPR. He invited the partners to feel free to change or edit things in any way they see fit.

Galya added that the team knows from the previous project the legal circumstances in particular countries (BG, NL, PL, UK, IT) but is not aware of the specific legal circumstances in AT, DE, FI, IE. So she invited the last-mentioned countries to provide information to reflect it in the Webscraping policy document if they have specific ones.

**3. Draft structure of the Methodological framework**

Galya explained the idea of the methodological framework. From her point of view, the methodological framework should be drawn up as a tool for statistical experts in NSIs to help them to implement a webscraping pipeline on enterprise characteristics in real statistical production following a harmonised methodology. While preparing it Galya read the manual for a traditional ICT survey and tried to reflect and transmit the specific features of Big data life cycle. The main goal at the present webex meeting was to discuss the proposed draft structure and then to define the final structure.

The draft structure is organised in nine main parts. The part relating to Enterprise characteristics reference architectures for BD production is also requested by WPF. Galya proposed first to discuss the main section of the structure and at later stage their contents. The final goal is for each section to describe the general approach, methodology, recommendations, instructions specific for enterprise characteristics data processing pipeline, which will be obtained during the project implementation phase.

Galya did not include the related software documentation since this will be organized using special repository provided by ESSNet project or WPC. She asked about opinions on the draft structure to check if this is the right direction.

Jacek confirmed it was the right direction, but pointed out use cases are not exposed and now they plan to be included in the annex of the methodological framework. . Galya considered we are at the implementation phase and the use cases are more appropriate for pilot projects. The use cases in the sense of implementation could be statistical variablefor the Big data pipeline. She will put the document on the WPC wiki section and requested for the feedback by the other countries by the next web-ex meeting in March. She also invited them for being proactive and add, delete or change something in the document.

4. AOB

Galya announced that the Review Board considered that it would be helpful to have a schedule when they can expect the draft versions of the WPC deliverables. Because of that, at the last CG webex meeting, Peter asked all WPs leaders to provide the date for each deliverable on which each WP plans to send the draft deliverables to the Review Board. Taking into account time needed for the Review Board to give feedback and then we to adjust the deliverable after the feedback received, Galya sent these dates to Peter, which are in fact drawn a month earlier than the deadlines the WPC team had.

Peter will make a draft overview for the Review Board to discuss and see if there need to be changes.

New dates for deliverables (sent to Peter)

1. ESS web-scraping policies – 27 May 2019
2. Methodological Framework V.1 – 24 September 2019
3. Methodological Framework V.2 – 24 September 2020
4. Experimental Statistics 2019, including reference metadata – 24 September 2019
5. Experimental Statistics 2020, including reference metadata – 24 September 2020
6. Starter Kit for NSIs V.1 – 27 April 2019
7. Starter Kit for NSIs V.2 – 12 October 2020
8. Quality template for statistical outputs – 30 June 2020

All agreed with the newly set deadlines.

Alex created the github organization essnetbigdata, available on this link: https://github.com/essnetbigdata . It's a github organization for the software source code for all Work packages and specifically for WPC. Alex started to fork repositories from the previous ESSnet. The repositories could just stay where they are right now and as the fork includes a link that is already helpful. Everyone who wants write access to the repo, could send Alex his/her github username, so he can invite you. At a later stage Alex should make a landing page, maybe teams/projects, etc., but right now he could already start collecting the software code.

After the official end of the web-ex meeting, Martin announced that from 1st March he is leaving ONS and Statistics and respectively WPC. He will provide Galya with the name of his replacement.

**Conclusions:**

1. Galya will update the description of use case 1 and send it to Olav for review.
2. It was accepted the general preference for less indicators with more different dimensions.
3. To have final draft description of use cases 1 and 2 and respective indicators by end of February and communicate them between all WPC countries to discuss it at the next WPC web-ex meeting.
4. Martin will finish the first draft of the web-scraping policy and will finish the use case 4 and descriptions of indicators for the Natural Language Clusters statistics before he leaves ONS.
5. Providing information on the specific legal circumstances reagarding web-scraping activities on enterprise characteristics in AT, DE, FI, IE by them.
6. All countries are welcome to make changes in the draft webscraping policy document.
7. Galya will upload the Draft structure of the Methodological framework on wiki so that feedback is received by the other countries by the next web-ex meeting in March.
8. All agreed with the newly set deadlines for the deliverables in accordance with the Review Board needs.

**The next webex meeting is planned for: 2019-03-19**