Minutes of the Pilots Kick-Off Meeting

DRAFT

Minutes of the kick-off meeting of the pilots track of the ESSnet Big Data II,

at Statistics Austria, in Vienna, on the 5th/6th of December 2018

Participants

Country	First Name	Second Name	Participant of WP
Austria	Magdalena	Six	WP K
Austria	**Alexander**	**Kowarik**	**WP K lead**
Norway	**Johan**	**Fosen**	**WPG lead**
Slovenia	Snezana	Vrhovac	WPG
Portugal	Sónia	Quaresma	WPK
Estonia	Arko	Kesküla	WPD
Estonia	Tõnu	Raitviir	WPD, WPI
Greece	Christina	Pierrakou	WPE, WPJ
Germany	Natalie	Rosenski	**WP L lead**, WPH, WPI
Belgium	Marc	Debusschere	WPA
Finland	Faiz	Alsuhail	WPA
Finland (LUKE)	Maria	Yli-Heikkilä	WPH
Slovakia	Boris	Frankovič	WPJ
Slovakia	Martina	Naňáková	WPJ
Germany - Hessen	Tobias	Gramlich	WPJ
France	Benjamin	Sakarovitch	WPI
France	Vianney	Costemalle	WPH
France - SSP	Bertrand	Ballet	WPH
Spain	**David**	**Salgado**	**WP I lead,** WPK
Netherlands	Piet	Daas	WP K
Eurostat	Fernando	Reis	
Poland	**Marek**	**Cierpiał-Wolan**	**WP J lead**
Poland	**Marek**	**Morze**	**WP H lead**
Romania	Marian	Necula	WPI
Bulgaria	Valentin	Chavdarov	WPB, WPG
Italy	Tiziana	Tuoto	WPK, WPI
Italy	Mascia	di Torrice	WPJ
Italy	Alessandra	Righi	WPG
Eurostat	Martina	Hahn	
Eurostat	Fabio	Ricciato	
Irleand	Aidan	Condron	WPI

Agenda

0 Welcome and Introduction

Alexander Kowarik welcomes all the participants of the Kick-off meeting of the Pilots Track of the Essnet Big Data II.

Marc Debusschere welcomes everybody in the name of Peter Struijs. He puts the pilot track in the context of the ESSnet and gives some background on the track and the set-up of the ESSnet Big Data II.

1 WPG Financial Transaction Data (Johan Fosen)

 Johan Fosen presented the planned actions for the WP on Financial transaction data (FTD), which also includes a part on Shared economy platforms. The work on FTD consists of two main parts. The first part is to get an overview of the FTD situation in each of the six participating countries, ranging from getting overview of the owner and other possible data providers, the relevant legislation for getting access, metadata, and the data's potential for contributing to different official statistics, the latter based on looking at accessed data if possible. The second part is empirical studies based on some of the promising statistics identified in phase 1, and the aim is to assess the potential of FTD in contributing to the official statistics, either by replacing the current data source(s), improving the current official statistics, or providing quality assessment of the current official statistics.

The Shared economy platform (SEP) part of the WP contains a part 1 and part 2 corresponding to the FTD part, but part 1 should be based on literature studies, and part 2 should only be started after discussions with Eurostat. In addition there is a part 0, being a literature study aiming at defining SEP in the context of official statistics.

Especially since part 1 of the WP involves some parallell activity that will clearly benefit from sharing and other ways of cooperation, Fosen presented also some thoughts about the different arenas for cooperation and how to use them fruitfully.

Discussion:

- The process of applying for access should start early in order to secure access in time to do empirical studies.
- Empirical studies could be started early – securing the presence of empirical studies in the WP is more important than to strictly follow the planned process of identifying promising statistics before starting an empirical study.
- Sufficient time should be allocated for editing, especially for the first deliverable (overview of data situation in each country) in order to avoid duplication of contents where each country more or less repeat what the others have already written.
- The National bank is a natural starting point for getting overview of the data situation in a country.

2 WPH Earch Observation (Marek Morze)

The WPH leader, Marek Morze, described briefly development of earth observation data market. He emphasized that the launch of the Sentinel missions from the Copernicus Programme opened a new chapter in applicability of remote sensing data ensuring free, continuous and open access to satellite imagery. He presented general information about WPH and more detailed information about needs, goals and anticipated results for each case study. He focused especially on the case study carried out by Statistical Office in Olsztyn (PL) as an example of prototype for data processing workflow and a plan for calculating area of specified crops using Object Based Image Classification and EO data.

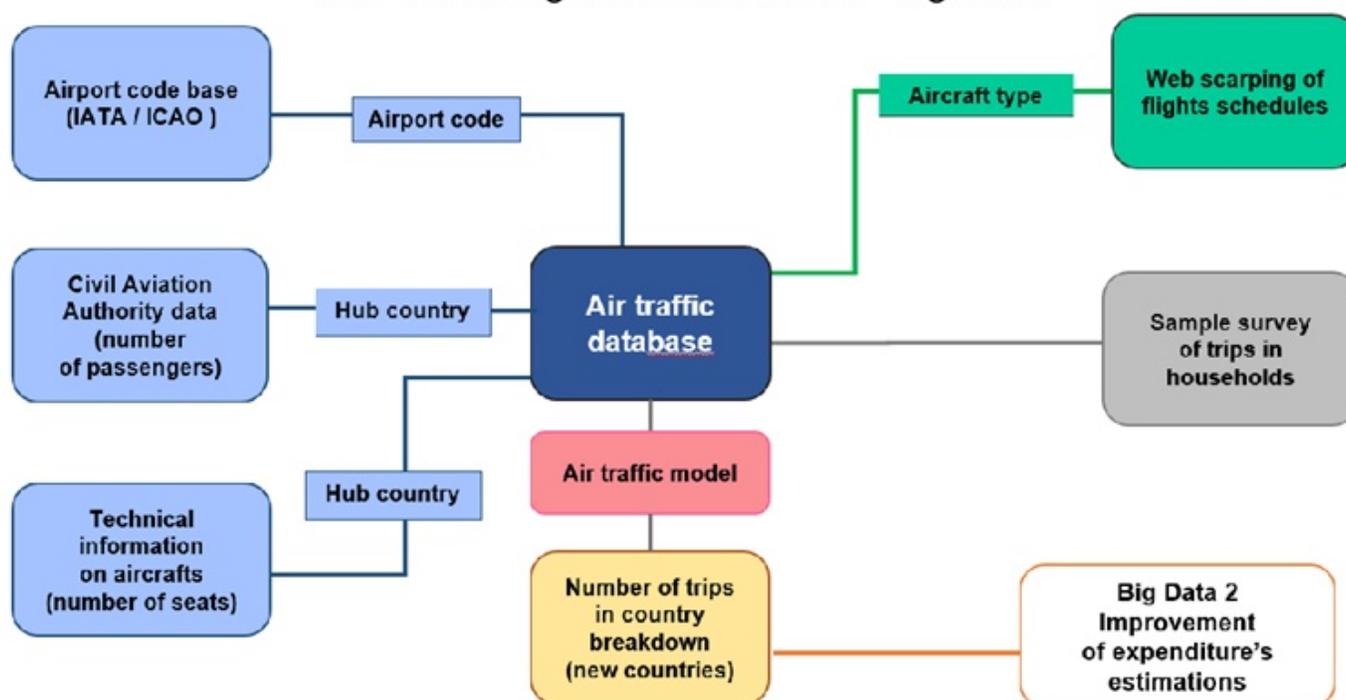# 3 Communication and Wiki (Marc Debusschere)

- The publishing of reports, meeting minutes, etc. will be done mainly by the wiki, which will be reused from the previous Essnet.
- Confluence will be used as a restricted site for confidential information.

4 WPJ Smart Tourism (Marek Cierpial-Wolan)

Marek Cierpiał-Wolan began the presentation with brief overview of the main concepts and objectives of the project. Next Marek focused on the method of verification of potential data sources, which in the future may contribute to improving the quality of tourism data, e.g. waste production data, as well as data on event related tourism. An important issue raised during the presentation was the method of combining big data with administrative registers and statistical databases that require the use of innovative statistical methods (Non Extensive Cross Entropy, Adjusted Spatial-Temporal Disaggregation Model, etc.). Marek stated that it is important to deliver data on expenditure by expenditure categories and average expenditure per trip in country breakdown. The potential relationship between the demand for offered accommodation throughout the year and the average price of accommodation has also been shown. Topic was ended with a description on how these estimates will be compared with existing statistics.

Next topic of presentation was the use of ESSnet Big Data I achievements in the WPJ package and the possibility of combining the results using the spatial-temporal disaggregation model. Short description of the model which will consist in combining the obtained data from web scraping of tourist accommodation establishments websites with database using adjusted spatial-temporal disaggregation method with regularization term was presented, as shown in the diagram below.

Data combining in air traffic model - Big Data 1

In the next part of the presentation, Marek introduced flash estimates of monthly data on tourists accommodated and nights spent at tourist accommodation establishments. He described flash estimates model containing bridge equations to derive monthly statistics from daily data from web scraping and time series model to obtain forecasts. Next he described how flash estimates will be compared with existing statistics.

The last part of the presentation shows the verification of direct tourism participation indicators in value added or GDP.

In the discussion it was pointed out that WPJ should take into account cross-cutting issues. It was mentioned that there might be necessity for closer cooperation not only with WPK and WPC but also with WPG – Financial transactions. Christophe Demunter from Eurostat put emphasise on the fact that one of the essential results of this package should be a prototype of a solid production system for tourism statistics. There were also a few questions about analysis of web scraping data as well as methodology prepared in presentation.

5 WPI Mobile Phone Data (David Salgado)

The presentation of WPI is accomplished in three steps: (a) Presentation of the research plan, (b) Brief discussion of common points with other WPs, and (c) Eurostat's presentation of the Reference Methodological Framework. The research plan is explained going from general working assumptions for our approach to the concrete partition into modular tasks to be carried in parallel to the extent feasible. Basically, eight tracks are proposed: (1) Access, (2) Synthetic Data, (3) Methodology, (4) IT, (5) Standards and Metadata, (6) Quality, (7) Application on Real Data, (8) Visualization. As potential common points with other WPs the following aspects are underlined: (a) the identification of the statistical problems to be solved (classical estimation problems in finite population, time-longitudinal studies, and network science problems), (b) the generation and use of synthetic data for many purposes (to avoid potential blockings due to data access restrictions, to test different data access scenarios, to test statistical models, to test quality indicators,...), (c) the representativity issue (inference models), (d) the construction of a common Reference Methodological Framework. Finally, Eurostat presents the main lines of this common RMF which WPI should feed according to the text of the project call.

6 WPK Methodology and Quality as cross-cutting issues (Alexander Kowarik, Piet Daas, Magdalena Six, Sonia Quaresma)

WPK focusses on three topics: Methodology, quality and typicfication. Each of these topics was presented seperately.

Quality:

Magdalena Six gave in her presentation on quality an overview of the quality-related work of WP8 from the last ESSnet,

but also an outlook what is planned in WP K of ESSnet II. The main quality related output of WP 8 of the ESSnet I was the deliverable "Report describing the quality aspects of Big Data for Official Statistics", in which the most important quality aspects were identified and described with respect to the work done in the other WPs.

Magdalena Six discussed the strenghts and weaknesses of the work so far: on the one hand, the report had input from the other WPs, who really worked with big data sources. Thereby quality aspects could indeed be described with respect to actual work with big data sources. On the other hand, the identified quality aspects did not fulfill the criteria of a quality framework, they were neither generic, nor exhaustive.

In the second part of the presentation, Magdalena focussed on the future plans. For the ESSnet II, WP K will provide quality guidelines and will try to identify quality indicators. WP K will create a template for quality reporting, which will also serve as the main tool to communicate with and extract information from the other WPs.

Again, there will be a literature review. Members of WPK will send out the template for reviewed papers to the WP leaders and ask them to provide information in a predefined form (template) about papers relevant for their work.

Typification

Sonia Quaresma presented the idea and motiviation behind the Big Data typification attempt, namely the big diversity of data sources, but the wish to use a standardised approach (the same treatments) whenever possible. If we have the possibility to typify a Big Data project or source, we could estimate the inverstment/resources needed for the exploration (and possibly the exploitation) of the source. The timeline/deliverables for the typficiation part were presented: the typification matrix for Big data projects and an evolution roadmap between the areas of the typification matrix.

Methodology

Piet Daas presented the third part of WPK about methodology. A methodology is a system of methods used in a particular area of study or 'a recipe'. He presented possible methodological issues identified in the previous ESSnet:

- Assessing accuracy
- What should our final product look like?
- Deal with spatial dimension
- Changes in data sources
- Machine learning in official statistics
- Data linkage
- Secure multi-party computation
- Inference
- Sampling
- Data process architecture
- Unit identification problem

The expected outcomes were presented with respect to methodology:

- Updated report on methodology from ESSnet Big Data I
- Methodological guidelines

It was pointed out that the quality and methodological frameworks and the typficiation are interacting and therefore need to be set up in a coordinated way.

Discussion:

- How should the interaction with the other WPs be organised, also with the WPs of the implementation track?
  → The quality reporting template will be introduced and discussed by a Webinar and a video
- Should there be a workshop for the WPs specifically focussing on quality,methods and typification?
  → Would be good, but no budget, rather concentrate in Midterm meeting on these aspects
- Should there exist seperate guidelines for quality and methodology?

7 Suggestions for Deliverables (Faiz Alsuhail)

Faiz Alsuhail was a member of the review board of ESSnet Big Data I and therfore read all the reports. Before we all start to write reports for the current project, he wanted to point out a few corner stones for writing good reports:

- think twice about the title
- keep it short
- keep your audience in mind
- focus on results (impact, relevance and meaning)
- give recommendations and mention bad practics beside good pratices

8 Big Data Test Infrastructure /ESTP Programme (Fernando Reis)

The Big Data Test Infrastructure (BDTI) was presented. It is an awesome based cloud service that will be set up according to the needs of specific use cases. One can easily select and use the latest open source projects on the cluster, including applications in the Apache Spark and Hadoop ecosystems. Software is installed and configured by Amazon. WPs from the ESSnets are possible use cases for the BDTI.

The second part of the presentation showed the curriculum of the ESTP programme for the upcoming year with a focus on the possible needs of the ESSnet, e.g. Big Data tools, AIS data, mobile network data and webscraping. Proposals from the ESSnet for new courses in 2020 are welcomed.

9 Parallel Discussion within WPs, Plenary with Wrap Up of the Internal Discussions, Conclusion

WPK

Participants: Sonia QUARESMA, Piet DAAS, Magdalena SIX, Tiziana TUOTO and Alexander KOWARIK

(David SALGADO and Jacek MAŚLANKOWSKI will be informed via Email and Webex)

■ Meetings

We first discussed briefly a plan for F2F meetings. There are to meetings foreseen dedicated only to WPK, one in mid 2019 and the second one in mid 2020, each 2 days.

The first meeting is planned in **May 2019 in Rome** and Tiziana will propose possible dates.

The mid-term pilots meeting is planned in M14, so December 2019 or maybe November 2019. Volunteering WPK partners are welcome to host the meeting, but it should also be possible to host at Statistics Austria again.

■ Quality guidelines

M9 first draft of guidelines

We try to move the second deliverable of the guidelines, the revised version from M13 to M18, so it can be aligned with the typification matrix.

Magdalena and Alex will prepare a proposal for the **structure of the quality guidelines until mid-February 2019,** shortly after (1-2 weeks), we will have the second webex meeting to discuss/fix this structure and possible divide the work load (or maybe in another webex, if the structure is not fixed easily). From **March until May we will try to fill as much content** as possible into the guidelines and then discuss it in Rome **and finalize until July.**

■ Literature Review

We will extend our literature review, this time **with focus on quality indicators/quantitative measures**. But also other papers directly connected to the content of the work packages can be included. We will use again the **same template** as in the last ESSnet. We already encouraged in the meeting the other participants of the pilots track to use the template and send us paper suggestion (to whom exactly should they send papers?).
It would be fantastic if again, Jacek is the main responsible person for the literature review.

■ Engaging other WPs

In September 2019 we will have a Webex meeting with all the Pilots WP-leaders (and possible a second one with all the Implementation WP-leaders invited). This webex meeting is meant as preparation for the Pilots midterm meeting. As a preparation for this meeting, we will have to prepare the following content (tentatively as a recorded video presentation):

-        **Presentation of the quality report template (Alex und Magdalena)** with a short tutorial how to apply it. We want to have two "filled-out" examples available **AIS (WPE Piet)** and **webscraping of enterprise characteristics (WPC Alex and Jacek)**

-        Presentation of questions concerning **methodology (Piet)**

- Presentation of questions concerning **typificiation (Sonia and Jacek), filled out example MNOs WPI David**, ready to be discussed in the May meeting

- Ask them to provide us with **literature and provide the template**

The Pilots midterm meeting could be arranged according our topics and each WP has to present methods&quality work, if anybody has an idea about that, please share with the group, but we will discuss that at future meetings.

We also discussed the possibility, to have an extra workshop, maybe back to back with the pilots (and maybe also the implementation) track midterm meeting which focusses explicitly on methods, quality and typification.

■ Next Meeting

We try to have a (short) Webex meeting in December, so that also Jacek and David are well informed and can make suggestions about the further way to go.

WPH

The WPH leader, Marek Morze, described briefly development of earth observation data market. He emphasized that the launch of the Sentinel missions from the Copernicus Programme opened a new chapter in applicability of remote sensing data ensuring free, continuous and open access to satellite imagery. He presented general information about WPH and more detailed information about needs, goals and anticipated results for each case study. He focused especially on the case study carried out by Statistical Office in Olsztyn (PL) as an example of prototype for data processing workflow and a plan for calculating area of specified crops using Object Based Image Classification and EO data.

WPI

During the meeting a first agreement is discussed to decide about the set of initial tasks to be carried out focusing on the responsible person, the participants, and the deadlines. This discussion is later on finished by email due to time restrictions and the high number of tasks identified.