ESSnet Big Data Pilots II

Implementation Track Kick Off

3rd - 4th December 2018, Wien, Austria

December 3rd 2018, 14:00-17:30

The meeting started at 9:30, as planned according to the agenda (see Annex 1).

Peter Struijs, in the role of project coordinator, asked Monica Scannapieco, responsible of the cross-cutting Workpackage on 'Process and Architecture', to chair the meeting.

M. Scannapieco asked for a round table presentation of participants (see Annex 2), then she briefly illustrated the agenda and gave the floor to P. Struijs for the first item:

Introduction to ESSnet Big Data II – Implementation Track

P.Struijs provided an overview of the results of the ESSnet Big Data I. Later, the main characteristics of the new project ESSnet Big Data II were presented, namely: duration (Nov 2018-Dec 2020), 28 partners (20 partners with reimbursement+3 NSI without reimbursement + 5 additional partners), three tracks (Implementation, Kickoff, Smart Statistics), organization in workpackages, review board, communication, financial administration. An high level overview of WPs in the implementation track was also described. Finally, P.Struijs underlined the opportunity of having a multi-beneficiary grant, instead of a set of mono-benificiary ones, thus remarking the importance of the European-level cooperation within the project.

Tomaz Speh, as coordinator of WPB – Online Job Vacancies presented the second item:

Overview WPB-Online Job Vacancies

T. Speh overviewed results achieved within WP1 Webscraping/Job Vacancies of the ESSnet. A specific focus on open challenges was provided, including: coverage, complexity of information extraction, problems on defining Online Job Vacancies for Official Statistics, differences at country level, etc. Later, T. Speh illustrated WPB work in terms of: tasks(Task 1 - Methodological Framework, Task 2 - Statistical output, Task 3 – Implementation requirements) and deliverables.

The role of CEDEFOP (European Centre for the Development of Vocational Training) was highlighted and a specific presentation by Vladimir Kvetan was introduced.

In the discussion, it was pointed out that WPB should take into account all the challenges identified during the pilot phase of the project but also be open for all opportunities that OJV data brings to the official statistics. Besides producing statistical indicators, OJV data has potential also in other statistical domains. For example in the field of classification revisions. It was particularly advised to create strong partnership and cooperation with CEDEFOP since their system is expected

to be finished by 2020 and its use in the field of official statistics could lessen burden for NSI considerably regarding acquisition and processing of OJV data.

Galya Stateva, as coordinator of WPC – Enterprise Characteristics presented the third item:

Overview WPC-Enterprise Characteristics

Galya first overviewed the results achieved within the previous project within WP2-Web scraping Enterprise characteristics. Then, she described the specific tasks of WPC, namely: Task 1- Web scraping policies, Task 2- Methodological framework, Task 3 -Experimental statistics (web scraping activity period March-April, two rounds 2019, 2020), Task 4- Starter kit for NSIs: selection of sw solution by use cases,Task 5-Quality template for statistical outputs.

Arko Kesküla, as coordinator of WPD – Smart Energy presented the fourth item:

Overview of WPD-Smart Energy

Arko started by describing the output and results of the previous ESSnet workpackage on Smart Meters. Then, he highlighted that participating countries have different starting points: Norway currently has data data aggregated at a monthly level, there can be linkability problems; Sweden does not have a real hub data yet and may have legal issues; Denmark and Estonia do have data, Denmark 4 million, Estonia 750 k. Arko also discussed the issue that EIC (Energy Identification Codes) are not interoperable and the highlighted the importance to interact with European Commision Smart Grid Task Force and to produce EU level deliverables (e.g. handbook).

The last item of the agenda, Overview of WPE-Tracking Ships was postponed to 4th December. The day was concluded by a discussion in which the following issues were highlighted:

- European dimensions within each workpackages. Each WP leader commented on the specific way in which the European dimension would be addressed. Among the discussed issues: the need for re-design definitions and processes if necessary, the need for having European-level solution for data access (like possibly the one provided by CEDEFOP), the need for shared software solutions (e.g. WPC Starter Kit for NSIs, WPD handbook), etc.
- Standardization of phases. The Big data life cycle proposed within WPC presentation is a starting point for sharing common phases. This is in scope of WPF.
- A question on success criteria was posed to WP leaders that answered by mainly highlighting (i) shared knowledge and resources (WPC); (ii) implementing solutions at EU level (WPB); (iii) reaching the outputs as defined by the WP (WPD and WPE).
- The strict interaction of WPK, in addition to WPF, with implementation WPs emerged. There was a suggestion of inviting the coordinator of WPK to the next overall meeting of the implementation task.

The work of the day was completed at 17:30.

December 4th 2018, 9:30-16:30

The second day started with the presentation of Anke Consten, as coordinator of WPE – Tracking Ships.

Overview of WPE-Tracking Ships

After a short introduction to the previous ESSnet, Anke detailed the tasks, namely: 1. Data access, 2. Definition, 3. Prototypes, 4. Experimental statistics, 5. Implementation and 6. Future perspective. It is relevant to mention on data access, the sources: EMSA, National AIS data of Dutch Inland Waterways and possibly World Wide AIS data (from ONS). Task 5 is on implementing AIS in statistical processes and could be a valuable input to WPF. Among future challenges, Anke mentioned: (i) getting access to data of EMSA, (ii) replacement of sandbox and (iii) retrieving Information on type and capacity of ships.

The subsequent presentation was by Monica Scannapieco, as coordinator of WPF- Process and Architecture.

Overview of WPF-Process and Architecture

After a short presentation of background concepts, Monica described the objectives of the WP, i.e. the definition of reference architectures necessary to carry out big data production both at national and European levels. The specific input to be provided to the WP by other WPs of the implementation tack were detailed in terms of:

Information architecture: Data (e.g. enterprises anagraphical infos), Structural metadata (e.g. data model of enterprises anagraphical infos), Process metadata (e.g. URL of the website where enterprises anagraphical infos where found), Quality metadata (e.g. accuracy of estimated economic activity), data building blocks

Application architecture: application services (e.g. text tokenizer), application building blocks

Solution architecture: specific IT solution adopted, in terms e.g. of building blocks and application services. Deployment scenarios, e.g. National level deployment vs ESS level deployment.

In terms of organization of the work of WPF with other implementation WPs, the following scheme was discussed:
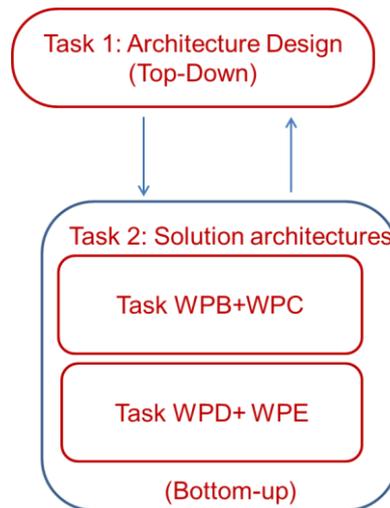
The figure shows two strands of work: a top-down one with draft artifacts to be revised and tested via a bottom-up one that takes into account specific solution architectures as implemented within implementation WPs.

The subsequent item of the agenda was the presentation by F. Alsuhail

Review Board Member Suggestions for deliverables - How to write reports that count?

Faiz outlined the following success criteria for good reports, e.g. good title, short reports, results first.

Later, Marc Debusschere illustrated some relevant point on communication:

ESSnet Communication

Main points illustrated:

- Web site under construction:
  https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page_new
- Contact email: essnetbigdata@economie.fgov.be
- Big Data for European Statistics conference (BDES 2020): To be held in Poland (city to be decided) in November 2020

In the subsequent discussion, the issue of having a github space for the project was discussed.

The meeting continued with parallel discussions carried out by each implementation WP and summarized in the following.

Notes on Parallel Discussion on WPB Online Job Vacancies

WPB participants discussed about the objectives and plans. Representatives from CEDEFOP and Eurostat also participated to the discussion. After a quick introduction, the discussion was mainly focused on the opportunities and interests of using the CEDEFOP data by WPB partners. All present partners expressed interest for using CEDEFOP data. It was agreed that CEDEFOP will

provide all necessary information and documentation for access to and use of their data as well as related documentation (description of available data and metadata) by the beginning of the next year. Countries should inform WP coordinator about persons who will access and use CEDEFOP data. The aim is to start the work on defining potential statistical indicators and other activities as soon as possible. The developments of prototypes already developed by participating countries will continue in order to be meaningful reused and integrated in the level of NSI as well as at the level of ESS.

Notes on Parallel Discussion on WPC Enterprise Characteristics

Present Participants:

- Alexander Kowarik (AT)
- Galya Stateva (BG)
- Kostadin Georgiev (BG)
- Jacek Maślankowski (PL)
- Martin Wood (UK)

Observers: Martina Hahn (Eurostat), Peter Struijs (NL)

Main topics of discussion

As WPC participants we had 1 hour parallel discussion session to really briefly discuss about the main objectives, distribution of tasks and timeschedule. All present partners actively participated in the discussion.

Task 1:

- Martin will start to work on ESS web-scraping policy using Netiquette (produced in the WP2, ESSnet on BD I) as a strating point;

Task 2:

2.1. Old Use cases (WP2, ESSnet on BD II) were updated and modified as follows:

- Use case 1: URLs Inventory
- Use case 2: Variables in the ICT usage in enterprise survey
- Use case 3: Validation of BR and NACE classification
- Use case 4: Experimental Language Statistics

At, BG and PL declared that they intend to work on the Use-cases 1, 2 and 3. The UK will define the new use-case 4 and will work futher on it.

As IT and NL did not participate in the discussion, they will be asked later if they agree with the so defined use-cases and if they agree, on which of them they would like to work.

By the end of January 2019 the description of all use-cases will be updated so that the new contents be reflected. Responsible participants as follows: Use case 1 – BG and NL? , Use case 2 – AT and PL, Use case 3 - IT?, Use case 4 – UK.

2.2. Defining of potential statistical indicators

The name and contents of the statistical indicators should be adapted to the updated use-cases as a starting point are already the defined indicators in WP2: Rate(s) of enterprises have websites, Rate(s) of enterprises doing ecommerce from enterprises websites, Rate(s) of enterprises that have job advertisements on their websites, Rate(s) of enterprises that are present on social media from their websites, Percentage of enterprises using Twitter for a specific purpose. All participants united around the idea that a clear description of the importance of the indicators is necessary, as well as of the way it is calculated.

2.3. Structure of the Methodological Guidelines

Galya presented shortly a draft proposal about the structure of the Methodological Guidelines, but due to a lack of time it was decided that it would be discussed at the next web-ex meeting.
The developments of prototypes already developed by participating countries will continue in order to be meaningful reused and integrated in the level of NSi as well as in the level of ESS.

Task 3

Part of the software means for webscraping used in WP2 are accessible in Git Hub and their use also in this project is recommended. For example BG will implement for web-scraping the BG and ISTAT software, PL also will use the ISTAT software. AT still considers which Software solutions for webscraping to use. IT and NL will be asked later for this. DE and FI will not carry out a real webscraping.
The aim is to start the work on defining potential statistical indicators and other activities as soon as possible. The developments of prototypes already developed by participating countries will continue in order to be meaningful reused and integrated in the level of NSi as well as in the level of ESS.

Task 5

AT expressed an opinion that it would be better for us to await the production of a result from WPK, which is responsible for the quality issues of Bigdata statistics. All participants agreed on this proposal. Due to this reason the presented deadlines and distribution of the subtasks in Galya's presentation drop out for the time being.
Other topics planned for discussion but due to time constraints will be postponed for discussion at the next web-ex meeting:
- How to achieve economies of scale
- Ideas for collaboration with WPB and WPJ regarding webscraping techniques
- F2F meeting – 2019 – when and where?

Notes on Parallel Discussion on WPD Smart Energy

The first topic addressed was: first deliverable in February. It was discussed that:

- WebEx meeting in December soon after the kickoff
- Exact structure of the report will be discussed in the meeting
- Writing the report will be done in Google docs for faster and easier communication between
- participating countries
- Ingergerd was willing to be the editor for the final version of the report

The second topic was: Data, quality and metadata. It was discussed that

- Every country describes the data what's in their database and also what they have added
- Quality of results - Comparison/validation with other data sources
- GSBPM 5.0 - which phases can be removed and which phases need to be added for smart meters project
- Quality management framework - the base will come from ESSnet BD1 (Report 2, table 32). We still
- need to do some modifications with it.
- Consult with WPK for improved quality framework

The third topic was: Second F2F meeting. The second face to face meeting (month 8) will be held in Estonia.

Notes on Parallel Discussion on WPE Tracking Ships

Participants:

- Anke Consten
- Eleni Bisioti

The discussion for this WP was very short. They addressed:

- Evaluation of the overlapping with the WP on smart meters, without reaching any specific conclusions on that.
- Discussion on the planning of WPF, with the outline of the the risk that the results of WP F are not ready in time for WPE.

Notes on WPF Process and Architecture

Present Participants:

- Monica Scannapieco (IT)
- Peter Stoltze (DK)
- Juliette Fourcot (FR)
- Yves-Laurent Bénichou (FR)
- Fischer, Bernhard (DE)
- Sonia Quaresma (PT)

- Paulussen, R.R. (Remco) (NL)
- Kostadin Georgiev (BG)

As observers: Martina Hahn (Eurostat), Peter Struijs (NL), Marc Debusschere (BG)

Main elements of discussion

- <u>WPF participation to implementation WPs</u>:
  - Monica Scannapieco (IT): Italy participates to WPC- Enterprise Characteristics and WPB-Online Job Vacancy
  - Peter Stoltze (DK): DK participates to WP D – Smart Meters
  - Juliette Fourcot Yves-Laurent Bénichou (FR): FR participates indirectly to WPB
  - Fischer, Bernhard (DE): DE participates to WPC- Enterprise Characteristics and WPB-Online Job Vacancy
  - Sonia Quaresma (PT): PT does not participate to any implementation WPs
  - Paulussen, R.R. (Remco) (NL): NL participates to WP E – Tracking Ships, WPC-Enterprise Characteristics
  - Kostadin Georgiev (BG): BG participates participates to WPC- Enterprise Characteristics and WPB-Online Job Vacancy

- Agreement on having virtual meetings monthly
  - **Action by Monica:** send out the schedule for 2019
- Recap of the tasks:
  - Task 1: Information and application Architecture Design (Top-Down)
  - Task 2: Solution Architecture Design (Bottom up)
    - Task 2.1 WPB+C
    - Task 2.2 WPD+E
- Task 1
  - Which modelling language to use in the WP work? The group agreed on using ArchiMate. Useful link: Archi software download
    - **Action by Monica:** send out a short intro to ArchiMate.
  - Activity to revise Big Data life cycle proposed within the previous ESSnet.
    - **Action by Sonia: send references to the work done within previous WP8**
    - The activity will be carried out by Sonia, Remco, Monica (with Jacek Manslakoswi from previous WP8 volunteering to help). Deadline: mid Jan 2019
- Task 2
  - Activity to prepare list of questions to be sent to Implementation WP leaders. The list of questions should be the starting point of the bottom up design and could be inspired by Monica's slide on Information and Application Architecture – what has to be provided by implementations WPs
    - The activity will be carried out by Remco and Berhnard. Deadline: mid Jan 2019
  - Activity to start modelling, according to the bottom up strand of work, architectural findings of ESSnet Big Data I.
    - The activity will be carried out by Juliette for WPB, Berhnard and Konstandin for WPC, Peter for WPD and Remco for WP E. Deadline: End Feb 2019

- **Other general elements discussed:**
  - Peter: ESS dimension to be always taken into account, think about evolution towards smart statistics, data governance as a topic to consider
  - Martina: need for having EU-level standards, requirements for Big Data Test Infrastructure, interaction with Eurostat EA Board
  - Standards: GAMSO should be considered

After the lunch break, the afternoon session started on time at 14:30:

Plenary Session with Reference Architectures Discussion

The first talk was by <u>Fabio Ricciato</u> from Eurostat. Fabio showed a Reference Methodological Framework for Mobile Phone Data. A focus on the different nature of such data was provided, showing CDR data and signaling data on one side as well as transactional and auxiliary data on the other side. Three dimensions were identified as basic ones to motivate the proposal of a Reference Methodological Framework, namely: Knowledge dimension, Institutional dimension and Temporal dimension (KIT as an acronym). Later Fabio introduced the *hourglass model*, i.e. a model including (from the bottom to the top):

- D-Layer, where D stands for data and embedding raw MNO Data
- C-Layer, where C stands for convergence and involving few common definition
- S-Layer, where S stands for statistical and including statistical indicators.

Each layer was subsequently detailed and it was discussed the importance of the C-Layer, and how it could be put in correspondence with the Information Architecture introduce by M. Scannapieco.

The second talk was by <u>Monica Scannapieco</u>. She briefly overviewed the main concepts of the ESS EARF (Enterprise Architecture Reference Framework) that are in scope of WPF work. Later she introduced a further standardization effort specifically thought for Big Data, i.e. NIST Big Data Reference Architecture, showing by example how the roles introduced by this reference architecture could be usefully mapped to the working context of the ESSnet project. Finally, she introduce a possible acronym for the work that WPF is going to produce, i.e. BIg Data REference Architecture (BIDREA) for Official Statistics.

The last talk of the session was by <u>Fernando Reis</u>. Fernando first introduced the Big Data Test Infrastructure (BDTI) as an effort by EC DIGIT. Fernando asks the possibility to have candidate use cases from the ESSnet to provide requirements for BDTI. There was a discussion on role of BDTI with respect to UN Global Platform for Big Data. It was discussed that a clear sustainability plan of this latter is lacking, and having a European level platform can be relevant for present and future investments. Fernando also illustrated the current ESTP courses and solicited for proposals of courses coming from project needs.

The second day of the meeting closed on time at 16:30, after final remarks and greetings by the project coordinator P.Struijs and the meeting chair M. Scannnapieco.9

Annex 1: Agenda

Location: ROOMZ Gasometer, https://www.roomz-hotels.com/de/roomz-vienna

3 December 2018

Afternoon 14:00-17:30: Introduction and WPs presentation (plenary)

14:00-14:30 Introduction to ESSnet Big Data II – Implementation Track (P.Struijs)

14:30-15:10 Overview WPB (T.Speh)

15:10-15:50 Overview WPC (G. Stateva)

15:50 - 16:10 Coffee break

16: 10-16:50 Overview WPD (A. Keskula)

16:50-17.30 Overview WPE (A. Consten)

Social Dinner

4 December 2018

Morning 9:30-13:00: WPs and Review Board presentation,Break up session

9:30-10:30: Overview WPF (M.Scannapieco)

10:30-11:00: Review Board Member Suggestions for deliverables (F. Alsuhail)

11:00-11:15 Coffee break

11:15-11:30 ESSnet Communication (M. Debusschere)

11:30-13:00 Parallel Discussion of WP Leaders with WP participants

13:00-14:30 Lunch Break

Afternoon 14:30-16:30: Plenary Session with Reference architectures Discussion

-Eurostat's Presentation on Reference Framework for Mobile Phone Data (F. Ricciato)

-Discussion on architectural issues (M. Scannapieco)

-Big Data Test Infrastructure and ESTP Courses (F. Reis)

Annex 2: Participants List

1. Peter Struijs (NL)
2. Marc Debusschere (BG)
3. Thomaz Speh (SI)
4. Annalisa Lucarelli (IT)
5. Vladimir Kvetan (CEDEFOP)
6. Črt Grahonja (SI)
7. Franis Saucy (CH)
8. Alberto Columbano (CH)
9. Jacek Manslakowski (PL)
10. Alexander Ciprianu (RO)
11. Ingerged Jansson (SE)
12. Jan Golubovič (LI)
13. Galya Stateva (BG)
14. Kostadin Georgiev (BG)
15. Arko Kesküla (EE)
16. Tõnu Raitviir (EE)
17. Johan Fosen (NO)
18. Anke Consten (NL)
19. Eleni Bisioti (GR)
20. Monica Scannapieco (IT)
21. Peter Stoltze (DK)
22. Juliette Fourcot (FR)
23. Yves-Laurent Bénichou (FR)
24. Fischer, Bernhard (DE)
25. Sonia Quaresma (PT)
26. Paulussen, R.R. (Remco) (NL)
27. Faiz Alsuhail (FI)
28. Martina Hahn (Eurostat)
29. Fernando Reis (Eurostat)
30. Fabio Ricciato (Eurostat)
31. Magdalena Six (AU)
32. Alexander Kowarick (AU)