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 1st F2F meeting WPC: Implementation – Enterprise Characteristics, ESSnet BD II – Sofia, 17-18 June 2019**

Participants

Aidan Condron - IE	✓	Johannes Gussenbauer - AT	✓
Albrecht Wirthmann - Eurostat	✓	Kostadin Georgiev - BG	✓
Dick Windmeijer - NL	✓	Luke Lorenzi - UK	✓
Donato Summa - IT	✓	Maia Papazova – BG	✓
Galya Stateva (PL WPC) - BG	✓	Olav ten Bosch - NL	✓
Jacek Maślankowski – PL	✓	Vera Ivanova – technical assistant - BG	✓

The representatives of FI and DE did not participate in the meeting.

Galya presented the participants and the proposed agenda.

The agenda is:

17 June 2019

1. State of the art of WPC activities and introduction to the meeting (G. Stateva)
2. Presentations by participants on the state of the work
   - Netherlands
   - Poland
   - Italy
   - Austria

18 June 2019

- UK
- Ireland
- Bulgaria

3. Future perspectives of Implementation – Enterprise Characteristics in the context of implementing Trusted Smart Statistics strategy (A. Wirthmann – Eurostat)
4. Methodological framework discussion (one country leading for the respective chapter)
5. Wrap-up and planning next steps

17 June 2019, Monday

1. **State of the art of WPC activities and introduction to the meeting (G. Stateva)**

Galya made a short presentation about what has already been done and what should be done by September – October this year. She reminded WPC is a continuation of the previous ESSnet project – WP2 and now WPC team are at the stage of trying to implement their results in real statistics production and in particular in ICT survey and also in business register. Galya outlined the main objectives of WPC: improve or update existing information (ICT usage in enterprises and BR); maximize the quality and quantity of the statistical outputs and to achieve important economies of scale: opportunities for sharing of resources at ESS-level. She announced the meeting objectives: overall view of the progress since last November till now; the present work done by WPC countries-participants; methodological framework discussion and planning next steps by October for producing the deliverables of documents and also to produce some experimental statistics since this is the middle of the project.

Galya presented the progress made from November 2018 till June 2019: draft ESS web scraping policies was developed by WPC members and commented at 2-3 web-ex meetings. She thanked all for their contribution. It was communicated with WPB and Eurostat, very fruitful feedback was received by Eurostat and reflected into the document, and in the end of May the Draft Policy sent to the Review Board and now the team is waiting for their feedback. By the end of June there should be a final version of the webscraping document after reflecting the Review Board feedback and Galya should send it to Peter and it should be published on the wiki website as a deliverable. It is a strategic document and it depends on Eurostat how it will proceed with it further.

The second task is the draft Methodological Framework – ver. 1 – the WPC partners had agreed on its structure and contents and the responsibilities had been shared and agreed among the countries. Also four use-cases and the statistical indicators had been defined. Galya and Kosta made a conceptual production model at the ESS level of WPC activities, it is a very first imagination how the business, statistical layer of WPC architecture should look like and it would be discussed at the Methodological Framework session.

The third task is producing experimental statistics by October (each country had already done webscraping on enterprise websites) and it will be published on the wiki page but the issue is how to present it to the Review board as a deliverable.

The last task was filling in WPF questionnaire in two versions and Galya said that Monica Scannaipieco welcomed the WPC work on the questionnaire and its share was complete and exhaustive.

Galya presented the use-cases: the first two could be directly implemented into the statistical (*URLs Inventory and Variables in the ICT usage in enterprise survey*) production process while use-cases 3 (*Data driven discovery of emergent enterprise classifications*) and 4 (*Enterprise Activity Clusters*) are proof-of-concepts.

Galya presented the deadlines for the tasks: the WPC deadlines for the Methodological framework would be decided the second day, it should be sent to the Review Board on 24th September and it should be published as a final deliverable on 30th October. The same deadlines are for the experimental statistics and on the following day it should be decided for the form to be presented, maybe a standardized template for each country.

Aidan shared his view about the situation in Ireland regarding the Business register. CSO receives the data from the Office of the Revenue Commissioner and they assign an ID number to the enterprises, so they use external administrative source. Kosta commented on the Bulgarian situation that unlike Ireland, we have Business register from the Revenue Agency but only with legal information with the same ID number; however we have the statistical information in SBR only for statistical purposes.

2. Presentations by participants on the state of the work

- **Netherlands**

Olav presented the Netherlands work on WPC. The work that CBS carried out is a mix of different projects: the main process for webscraping websites, URL finding and retrieving enterprise characteristics from social media. There is also generic software developed for these projects. The combination of google search and scraping websites is a iterative process – finding sites and webscraping them, webscraping and finding sites. CBS used URLs from the Chamber of Commerce. They derive the URL using the Google custom search API and they determine which of the results to assign to an enterprise. For each search query CBS uses the first 10 Google results. The work is a continuation of the work in WP2 which was based on research of ISTAT. CBS has a General Business register (GBR) which contains URL, but only for about 1/3 of the enterprises and of course CBS does not know for sure if the registered URL is correct. CBS uses machine learning (support vector machine) to decide on whether a URL has top be assigned to an enterprise based on a training/test set of 3300 enterprises taken from the enterprises in the GBR for which a URL is known and additional data from a scraping company: data provider. CBS performs multiple search queries on Google and analyses the first 10 results for each search query. There are about 30 features defined on all Google search results. The snippet (a small text returned on some search result) is important for that. CBS also uses the Google PageMap that is returned with each search result. This is a structured JSON description of the search result if micro format is available on the web page.

Olav explained the feature importance in URL finding: eqTitleLegalName_max: maximum Jaro Winkler (JW) distance of enterprise legal name and Google title; Zscore: how prominent the candidate URL appears in all search results (the probability the result to be the correct one); Sequence score: average rank of candidate URL in all search results (the best score); the correlation between the different features. Dick explained the correlation between predicted and search results and presented the tests - ROC curve and scores per label. Olav then explained the generic scraping process and variables retrieved. They start from the postal codes. There are a lot of domains for enterprises found.

Olav also presented the results of another project – Enterprise characteristics from social media. This is some exploratory research to the possibilities of the use of social media for statistics on enterprises. It was based on messages collected by www.coosto.com from 2017-06-26 to 2017-08-04. CBS explored the correlation between the social media database and ICT survey. There were much more messages in the social media than in the survey.

Olav mentioned the existence of some generic open source software - S4SGoogleSearch Node + a Python version is in use, not yet published and generic scraping software in Python: Python, Postgress, parallel scraping, storing webpages as the work is in progress.

Olav summarized that webscraping for all enterprises has been set up, many variables were derived; URL finding based on earlier research using Google API and ML is getting mature; CBS want to combine both and do maintenance of domains via Google URLfinding. Social media as a source is also promising. Other work mentioned is NACE detection (Kuhnemann, ms thesis).

Galya asked Olav if they are intending to do webscraping for small enterprises (below 10 employees). He said that these did not have the focs yet, but that may be the case in the near future.Aidan commented that in Ireland there are 400 000 registered enterprises but 80% are sole persons – farmers and taxi drivers and don't have websites. Donato shared the Italian situation that there is a law that each company should put their fiscal code on their home page. All the participants shared their thoughts and their countries' specifics about the difficulties in finding and verifying URL so that to be most effective and least expensive. Galya requested sharing of the Dutch software on Github. Now there is IT, BG and PL software. Albrecht recommended using the same language in doing webscraping and sharing the results. All agreed that results even within a country differ from various webscraping activity even for one and same URL.

- **Poland**

Jacek presented the Polish work on WPC. He outlined the Polish experience on webscraping, the changes in their approach since WP2 and made a summary and noted the steps for future work. He started with the technical aspects of the case study: the used tools are Apache Spark, the language is Python with pyspark and the results of processing are CSV files. Like NL Statistics Poland uses 4 URL. They did webscraping in March-April with the enterprises participating in this year's survey and don't have the results from ICT survey yet, so the comparison could be done at the end of the year. Statistics Poland don't collect e-mails and it is very difficult to identify enterprises by email. They use URL verified from the BR, URL from email, but some emails belong to provider, not enterprise. They also scrap websites and got URLs. This is the sample and it comprises enterprises of more than 10 employees. According to Jacek the sample is too big. They don't scrap the total size of the population. Some websites don't respond. Webscraping was accomplished in March/April 2019 - 6 times every 1-2 weeks. The data are very similar to data received in WP2.

The changes since WP2 are vulnerabilities (PDF, no-html files) but now there is response within 5 seconds, 60 seconds to download the webpage. There are still 30% missing enterprises of the population. There will be improvements. The future work will be aimed at other attributes of enterprise, the use of the machine learning, extending the population and supplement of non-response websites.

Aidan asked what portion of sectors were presented. Jacek clarified it was the sample of ICT survey but he would check that since he has access to the database and would see if there is under coverage or over coverage of some sectors. Jacek added that from 6 million enterprises only 200 000 enterprises in the BR have websites. Like other countries they also have question about www. and email.

Galya pointed out that the first indicator is rate of enterprises having websites measured by the ratio between the number of enterprises having websites towards the whole number of enterprises in the reference population in ICT survey. The aim is to standardize all countries. Jacek suggested that they could repeat this step for the population but they have too much population. Galya stressed that ICT survey published estimates for whole population, even estimates for webscraped data. Albrecht agreed with Galya that the same base for calculation of the indicator is necessary for comparison between the countries. Galya added that BG ICT department uses the webscraped data to improve their data because in BG especially there is a problem with declaring having websites due to lack of willingness by the respondents to fill in more data in the questionnaire. In BG SBR only 2000 enterprises declare having websites but BG WPC team found 11 000 websites.

- **Italy**

Donato presented the Italian work on WPC - Data driven discovery of emergent enterprise classifications. He recapped the use case description, use case objectives and focused on the most important – the working process. Use-case 3 is an experimental one. Given a huge text corpus obtained by a massive web scraping activity on enterprise websites, natural language processing techniques (like, e.g. Word Embeddings, LDA, etc.) will be used to: 1) explore the semantic content of the corpus; 2) identify meaningful semantic and syntactic patterns; 3) use the discovered patterns to define data driven enterprise classification and 4) associate each enterprise with the most appropriate class of the discovered classifications. The use-case objectives are to obtain one or more new data-driven enterprise classifications and associate each enterprise with the most appropriate classification (find the corresponding distributions of the scraped enterprises).

Donato firstly outlined the steps of the working process shortly and then in detail: 1) URLs acquisition with an output of list of urls; 2) URLs scraping – output scraped text; 3) scraped text file cleaning with a cleaned textfile; 4) Word 2 Vec algorithm with an output trained model; 5) using WordEmBox for visual inspection and finally the result should be 6) clustering algorithms with clusters. In the first step URLs acquisition from the data in the beginning of the process there is a list of names and ids of enterprises and at the end via the program URLScorer there are scores of each URL which show the probability of correctness of these URLs – the higher the score, the greater is the probability this URL to be official. So at the end of the first stage there is a list of official URLs which is input for the next stage – URLs scraping. It goes through the Java programme RootJuice and the resulting scraped content is delivered in a CSV file format. ISTAT uses the same webscraping tools as in the previous project and they are free and accessible on the Github.

Galya mentioned that BNSI also used ISTAT software and got the same results but ISTAT lacks the tool for analyzing in order to keep up with BNSI. Donato further explained the next step scraped text file cleaning done with custom Python program for cleaning operations. Then is used the algorithm Word 2 Vec which is a word embedding model that maps words to vectors of a vector space – similar words appear in similar contexts. ISTAT are interested in predicting the words in a given context. They are dealing with a huge text corpus. Word Embedding models are data-driven. Word2vec implements a neural network to predict Y given X. Two mode for word2vec: the model predicts the context given a word and the other one - the model predicts a word given the context.

Donato explained the exploring and visualizing big embedding models through graphs. The objective is to explore the model around a semantic area and to represent the relationships between words emerging. He also explained the geometric graph. Then he continued with the 5th step - WordEmBox is a web based application. It can be used for word embedding trained model exploration. There are 3 tools available: Graph, Affinity and Word analogy. The ISTAT tool (graph in particular) could be used for: visual analysis of the data, hypothesis confirmation, knowledge discovery, relationships discovery and social network analysis. Finally clustering algorithms could have 2 possible strategies: K-means (on the whole space) or dimensionality reduction (PCA or TSNE) and then K-means.

Johannes asked if information about the officials in the enterprises is collected in this way. Donato clarified that preselection of pages only for the content of the top page is done in order the noise to be removed. The algorithm decides which words are more important and ignores words with smaller or bigger length. Galya commented that ISTAT are trying to generate an enterprise classification but not exactly the NACE one. Donato confirmed that the goal is to produce one maybe close to or different from NACE or maybe even to extend NACE. Aidan suggested that ISTAT make time series so that they could trace the changes during the years. Donato said they are using a sample and there will be algorithms very soon. The relationships are automatically generated and there are no guidebooks but only intuition and personal experience to get the best possible configuration. Galya announced that by now only ISTAT has worked on that use-case but PL and AT also have interest to work on it. AT confirmed their willingness.

- **Austria**

Johannes presented the Austrian work on WPC - URL-Validation & Social Media presence. He talked about the setup – the webscrapping server is outside of STAT Firewall; the search for URLs is done with URL-Searcher from ISTAT / BING-Search (Yellow Pages API) and the webscrapping and textparsing is done with R. The first 10 Hits from ISTAT URL-Searcher are taken. There are blacklisted URLs and the blacklisted URLs are excluded from BING-Search results. For the ICT there is a sample of 5200 and the results from BING-Search considering blacklisted URLs are 4690. Then they scrape URL using Selenium (R-package RSelenium). There is a 6-step procedure which is repeated for each enterprise and URL. The next stage is URL validation: loading scrapping results and parsing them using R-Package rvest (libxml2 C-library); trying to identify enterprise in scrapped text; according to the Austrian Media Act enterprises must state VAT and or company register number (CRN) as well as name, address,. . . on URL
Link to this information on main page and any sublink. URL is defined valid if at least one of the 2 informations is found: VAT and or CRN or full name and address as listed in AT register.

Johannes cited data for number of valid URLs, URL valid by VAT/CRN and or name, address,. . . and also gave data for comparing webscraping with ICT. But it is a validation issue that not all companies adhere to Austrian Media Act. He commented the case with the searched social media presence and the issues raised - quality of response variable questionable - link to social media platform on homepage - survey response indicates no social media usage; reference to social media profile which is not owned by the enterprise - How to check the social media profile?

Galya noted that BNSI developed their own software for social media but BNSI use the Polish one. She asked Johannes if they searched for OJV also, but he answered they didn’t. But Statistics Austria did also e-commerce search. Albrecht stressed on the importance of the data what is presented on the social media. Jacek suggested restructuring and refreshing of the questions in the ICT survey questionnaire.

18 June 2019, Tuesday

- **UK**

Luke presented the work done on WPC - unsupervised classification of businesses from website descriptions by the Office for National Statistics of the United Kingdom. He gave the motivation, introduced the pipeline so far, gave details about each step in the pipeline and outlined the current challenges and the next steps. As regards the motivation, in UK Standard Industrial Classifications (SIC) are used by statistics agencies to categorise businesses by economic activity. For the reasons of difficulty to keep up to date - there are around 4.1 million registered companies in the UK and the changing structure of the economy, it means that SIC will constantly lag reality; therefore, it will be unable to quickly identify new and emerging industries & others declining in importance. So the ONS decided to apply Data Science and Statistical Techniques to see if these limitations can be overcome.

Luke presented each part of the pipeline in detail. The first component of the pipeline is the data - Companies House is the registrar of companies in the UK. All registered limited companies must submit annual financial statements, which are publicly available. Businesses must report an SIC when they first register with Companies House. ONS can link the business descriptions to these SIC codes. Novabase (a 3rd party company) used the Google API to find the most likely candidate for each of the 4.1 million business on Companies House. It scraped the homepage and followed specific links if present e.g. “About Us”. It assigned a trust score for how certain they were that they got the right business.

The second step in pipeline is the preprocessing: removing descriptions with a very low trust score (≤ 0.4). They also remove short records, (< 30 characters) as these are unlikely to include useful information. Next step is using Models: LDA and Doc2Vec. LDA uses a Bag of Words feature representation to represent a document and uses the pyLDAvis library in Python. Using Doc2Vec ONS are able to represent documents as n-dimensional vectors. The vectors of truly similar documents should be similar to each other – allows the statisticians to cluster in the embedding space. They could also use the elements of the document vectors as features for a classifier to predict SIC. Then in the pipeline comes the Dimensionality Reduction where clustering in high dimensional spaces can be challenging; distance can become less meaningful and the data can become more uniformly distributed in the space.

Principal Component Analysis (PCA) is used to reduce the dimensionality of the document vectors. Its advantages: very fast and more scalable compared to methods like T-SNE; its disadvantages: unable to capture non-linear relationships within the embedding space. In the clustering stage clustering approaches can be applied to both LDA and Doc2Vec outputs. ONS have chosen HDBSCAN as their preferred clustering algorithm - a hierarchical, density based approach to clustering. Unlike algorithms such as K-means, they do not have to ‘guess’ how many clusters are present in the space. Using HDBSCAN, observations can be left unassigned to a cluster. Then there are two ways for validating the received clusters: internal measures – separation; compactness, etc.; and external measures - Do the clusters of documents agree with the SIC codes?

Finally Luke outlined the current challenges: it’s difficult to determine how well Doc2Vec is fitting to the data without a large amount of user input/inspection. PCA doesn’t seem to reduce dimensionality by any useful amount e.g. for Doc2Vec vectors in 100 dimensions, about the first 70-80 principle components are still needed to explain 90% of the variance. He concluded with the next steps: establishing metrics for cluster performance (internal and external); pipeline improvements - different word embeddings and other approaches to clustering; and exploring bias/coverage in the dataset.

Olav asked which model is better and Luke answered that LDA is better than Doc2Vec in terms it seems easier to handle. Olav also wanted to know if that was a one-time exercise with that company and the answer was yes. Albrecht asked about comparison of the models in validation and Luke answered it was done. Johannes commented about other approaches. Galya asked Luke to share on Github.

- **Ireland**

Aidan presented the work of CSO on WPC. They started this project 8 months ago. CSO used the deliverables of the first phase. CSO is interested in URL retrieval, e-commerce, social media and NACE classification. They use the concepts of generic and specific webscraping. CSO webscrape the information from the main page with contact details. They use the emails, but many of them are blocked if they are not in the CSO list – like gmail, etc. They have access to 3 data sources. They apply automated web searches + deterministic & probabilistic methods used by ESSnet partners. So far CSO implemented one deterministic linking method – e.g. matching single occurrence email domains. Now they are working on probabilistic matching drawing on Italian method. They use deterministic and probabilistic match and this is familiar to Donato because it is the Italian scheme for linking data. But there are many "dead" pages with no information. LDA method was used to fit a topic model to the data.

In these tests a sample of 20,000 websites was used and the value of K was increased for each test run. Aidan talked about e-commerce watch – similar to ISTAT and current social media detection. There is breakdown of different social platforms. The second stage is scraping: ESSnet generic vs specific scraping. CSO use generic one – cache frontpage text and links. The data are captured in JSON and flattened to CSV. At this stage trying probabilistic approach is too expensive. CSO will have some experimental statistics.

CSO have data protection concerns with webscraping. So they should not share their business register with search engines. When the data officer reviewed webscraped data, it was considered there are no serious data protection problem. The GDPR exclusions for statistical purposes should be considered for lawful use of information. They could be used for webscraping. NSIs have their own Acts but GDPR should be used as a common approach for all NSIs Acts.

Aidan explained the institutional and politically conservative context for the strict adhering to data protection with GDPR. Kosta and Galya clarified that in BG the information from the BR of the Revenue Agency is public. Aidan said that open data is not all data made public. Albrecht asked about finding only companies' websites and scanning only for variables for statistical purposes and then there would not be data protection issue. Aidan replied that still there would be problems with the websites of sole persons who publish personal information on the websites. Albrecht suggested a set of recommendations and conditions for proper webscraping to be developed and Galya accepted the idea and told the partners that maybe a separate chapter or a subchapter in the Methodological framework should be added.

- **Bulgaria**

Kosta presented BNSI progress on WPC. BNSI works on Use case 1: URLs Inventory and Use-case 2: Variables in the ICT usage in enterprise survey: enterprises engaged in web sales on their website; enterprises that are present on social media; enterprises using Twitter for a specific purpose and enterprises having specific features of the website, e.g. job advs. BNSI use their own software (PHP), ISTAT software (JAVA, Solr, R) and Statistics Poland (Python) (software for social media presence). BNSI use initial data from SBR.

Kosta presented the architecture of URLs Inventory for the stages collection, processing and analyzing. Galya stressed that on recommendation by Monica BNSI tried to be compliant with WPF in all implementation packages – same architecture, artefacts, language. Kosta showed the experimental results – output table by NUTS region results published on BNSI Intranet website. He explained the results of the indicators of URLs Inventory and the variables in the ICT usage in enterprise survey. Galya pointed out that BNSI will continue to use the Polish software as most successful. Kosta finished his presentation with the results of ICT usage in enterprise with less than 10 employees – BNSI are in the beginning of such statistics. BNSI are planning to use ISTAT software to webscrape the small enterprises and with machine learning.

Maya Papazova explained the search and comparison of enterprises which own e-commerce. Now there are the results of 2018. Galya added about the comparison between the webscraped data and ICT survey and the expected results in September. Last year BNSI scraped data for OJV but only a small percentage of enterprises published them on their websites. Maya explained that there are even enterprises that do not use computers, e.g. they use outsource counting. Maybe the enterprises do not understand correctly the questions.

Galya shared her thoughts that Python and ML is quite new for BNSI but she is optimistic for its development. She suggested that Eurostat should have the initiative to organize such training courses, e.g. ESTP. Galya has a proposal to the BNSI Director General for training of 10 people in ML methods – IT experts and methodologists. Maya said that 3-day ESTP courses are not enough. Albrecht agreed with the idea for course next year, probably an online course. All greeted his suggestion. Dick announced the existence of free online courses in Dutch universities in English and he promised to send the link. Galya shared her opinion that the support from the top management is of key importance. She also said that now the NSIs meet with the competition of private companies in webscraping.

3. Future perspectives of Implementation – Enterprise Characteristics in the context of implementing Trusted Smart Statistics strategy (A. Wirthmann – Eurostat)

Albrecht presented the enterprise characteristics in the context of implementing Trusted Smart Statistics strategy – a document discussed at ESSC. Currently many Directors General have undertaken actions and initiatives to access to privately held data for statistical purposes. This approach is more relevant to those issues about dealing with private companies' data. He outlined the main principles of trusted smart statistics set in the document. It's a multi-source statistics based on multi–purpose data sources while not internalizing all data and pushing computation out for getting real interesting data with using modular methodological frameworks for coping with technological developments and using data without sharing.

There are intermediate layers producing intermediate data. The additional sources should be integrated to produce statistics. It requires changes in production statistics. It is principally one survey for one purpose. Production is not centralized in one department, there should be a horizontal organization which is to organize statistical production process. There is relation with GSBPM – when there is information need, sources are found and the data collection model is designed. But this statistical process might not fit into our statistical world – instead of designing the sources, information extraction models are designed. For this knowledge extraction a framework is introduced called Hourglass model and it is tried to be implemented in mobile network data where there are different technologies – 3G, 4G, 5G, quite heterogeneous data. There are problems when technologies are changing. The technological problems should be separated from the statistical ones. Convergence layer with common definitions is determined. There are raw microdata at the MNO Data D-Layer, standardized /uniformed microdata at the Convergence C-layer and aggregates, macrodata at the statistics S-Layer. All processes – collection, processing and dissemination of data – are in the control of the NSOs.

In Big data there are some issues – first, the volume of the data does not allow wholly internalizing the process and other issue is privacy. There is a risk of attack. Pushing computation out minimizes the risk. The data are not internalized for the sake of having them but for using the data for extracting statistical information. For the future the aim is instead of pulling the data in and processing it, the data should be preprocessed and pushing computation out gives trusted smart statistics. Instead of sharing the input data, input data would be used for a certain purpose. Nowadays or in the past NSIs have agreements to use input data within NSIs. In the future NSIs will probably agree on the query and only pulling in the results of the query and there will be less subjects of personal protection. The question is if the companies will be willing to do that and the additional resource for setting queries. So these are the principles Eurostat want to develop for the future especially for mobile network data. For each data source it could be written special principles.

Albrecht listed the priority areas: trusted smart surveys; mobile network operator data for human presence and mobility; transport and logistics data; statistics from smart systems (incl. energy); Earth observation; inferences from data on the web. For the trusted smart statistics there were hackathons and some data were collected – data protection issues are important. For the mobile networks a unified methodological view was developed - Reference Methodological Framework [RMF] for processing MNO data for Official Statistics. The aim is to facilitate interworking MNO-ESS at technical & organisational level, to ensure consistency, reproducibility, evaluability and portability of processing methods (between MNOs and SOs), to provide concrete basis to clarify legal aspects (GDPR) and to enable multi-MNO analysis (fusion of data from different MNO).

For transport and logistics there is use of tracking data to provide long-distance transportation and logistics. The initial focus is on ship position data and there is extension to air and railway traffic data; beside there are flash estimates of economic indicators. The smart devices ensure proofs of concept and pilots - smart farming, smart traffic and IoT for smart cities represent trusted statistics from smart systems.

As regards inferences from data on the web the European Commission is interested in online job vacancies, there is link also on CEDEFOP for skills, job vacancies. There is also an internal project in Eurostat for job vacancies for EuroGroups Register, a small prototype for Wikipedia for turnover of enterprises and also for such enterprises with a seat outside Europe.

Eurostat are developing Web Intelligence Hub (WIH) – they have interest in online job vacancies and have an agreement with CEDEFOP to share progress. The scope of the Web Intelligence Hub also is: enterprise; Wikipedia (EuroGroups Register); linking to existing statistical infrastructure - linking to business register. Albrecht explained the services to be implemented by the WIH - providing support to ESS partners in: data acquisition (web scrapping, APIs), trans-national data agreements, partnership models for national data agreements, IT infrastructure and tools, analytical services (e.g. NLP), methodology, regulatory aspects, skills (training material), R&D collaboration and governance.

The principles of WIH are: ESS hub; serving national and European needs; modular structure; defined processes and products to be guaranteed; priority to working together, possibility to act individually; programs should be open source; transparency as much as possible; common used processes should be certified and audible; lineage of data and processes and intermediate products usable by all partners.

By the end of the year Eurostat would like to make a Memorandum of Understanding with CEDEFOP about the OJV and start finding appropriate architecture and in 2020 start to build the components to mirror this CEDEFOP system in Eurostat infrastructure. Besides Eurostat will use the output of ESSnet, the methodological framework, not to be based on peer-to-peer base but on a general approach. Then in 2021 a joint production system should be set.

Regarding webscraping policy it is foreseen a discussion at TSS task force, at the methodology working group, at Directors‘ groups, PG and ESSC. It will depend on the view of the ESS what will be the status of this document.

Albrecht outlined the text specified in the project application of WPC and stressed it could be relevant for WIH. He recommended publication of experimental statistics on a project wiki platform and on the dedicated page of Eurostat. Experimental statistics should provide reference metadata based on a standardized metadata structure (e.g ESMS). Starterkit for NSIs for web scraping on enterprise characteristics should consist of procedures for testing and maintenance of web scraping. It will use solution architectures of WPC shared to the WPF for design of a data and application architecture for big data production. Albrecht said a quality management template for web scraped enterprise characteristics should be developed and the results of previous ESSnet should be further developed. The follow-up questions raised by Albrecht are: How to communicate the results among ESS? How to enable NSIs to use the results? How to ensure sustainability of results? Support to NSIs not participating in the WP should be given. Could web intelligence center contribute to communication, training, support?

The discussed questions concerned trustworthiness of smart statistics, the perspectiveness of WIH, strategic character of Webscraping policy, BR and OJV and respective architecture.

4. Methodological framework (discussion session)

Galya started the methodological framework discussion after she stated her views about the document. Initially there is the generally agreed structure of the document but the last month she tried to change its direction a little inspired by the work of WPF, because they are the "umbrella" of all implementation packages and they try to standardize the language, artefacts, solution architecture and business layer of the architecture. WPC team should prepare guidelines for the statisticians to implement in their production process. In WPC there is another document – Starterkit intended to be a guide document for developers. For that reason she slightly updated the name of the document – Reference methodological framework for processing online based enterprise characteristics (OBEC) data for official statistics v.1. This is a new concept – "online based enterprise characteristics" and it is an output of the discussions at the Rome meeting in May during WPF physical meeting and also this concept derived as an output from WPC questionnaire. She asked all participants if they agreed with that idea for the document.

Olav asked if WPF had prepared a document. Galya clarified that there are still draft minutes of the meeting in Rome and besides she sent everybody their document – Big data life cycle. She followed the Big data life cycle – 4 phases: collection, processing, analyzing and dissemination. She put the first question: Do you consider this document could be used by statistical experts or not? She means when implementation is concerned that statistical experts should be able easily to work with the new data. In second chapter Business context the main concepts are defined, also population, observation variables and statistical indicators are defined and how they are calculated – all this is statistical methodology, even for big data.

Galya further explained that she started from the ICT survey manual and tried to transfer it to BD sources. But it's not possible to refer to ICT because doing the processes is different and WPC team can only refer to some parts of ICT survey. Galya doubt if there is necessity of the presence of the Legal aspects of webscraping on enterprise characteristics in the document. She suggested its only referring to the policy and deleting this chapter and the first chapter to be Business context.

Albrecht advised on having here the general methodologies reflected in the document and starting with the characteristics of data and specific processing and methodologies and methods that exist. Galya agreed for the need of a specific chapter for more general explanation of Internet data. Albrecht pointed out that population is different with Internet than the survey of enterprises, also the statistical units are different. He recommended starting with general concepts describing the differences, WPC have statistical enterprises and their indicator is the information for the economic activity on enterprise website. Olav mentioned there are different kinds of websites – citizen or community websites. Galya noted that the main difference is that the statistical indicators produced by ICT survey have clear quality stamp for official statistics but WPC indicators at the moment are not with the same quality stamp as official statistics and hopefully they are to become such quality. Albrecht pointed out there are different approaches and methods for delivering quality statistics.

Galya said that the population is the population of ICT. Albrecht pointed out that 40% websites are from enterprises, 50% are from social network websites and they arrive at statistics with webscraping based on this. There are considerations for the statistical units not going directly to enterprises. There should not be made a difference between populations. Olav stated that there is an opportunity for mapping of different populations.

Galya suggested maybe it would be better to keep the previous name of the document – just Reference methodological framework. Olav shared his opinion that there is not a big difference but the terminology should be kept in line with the other WPs and WPF. Galya explained that WPF will have their v.1 at the end of July and there is an issue how to coordinate with them because WPC expect output from WPF. Albrecht advised to do Methodological Framework v.1, then after receiving WPF v.1 to be clarified the issue with terminology and it to be added. Olav suggested having more generic approach.

Then the conceptual production model prepared by Kosta and Galya was discussed and amended according the team’s views. Each concept should be described with a text. On agreeing with this conceptual production model then it would be possible to prepare the generic version of the methodological framework because it’s a fundamental understanding of the processes. It was agreed that there are 5 actors as data providers. According to Albrecht the data provider is realization of the data source. Galya stated that for her data source is Internet. Albrecht clarified that besides Internet, there are Yellow pages and statistical register as more general. He considers enterprises’ websites and search engines a data source. According to Albrecht part of WPF should be to define actors generally and then to apply within a project. Galya answered they are defined generally but each package should specify them. WPF has already defined BREAL architecture for all implementation packages. Now it is WPC turn to define their specific conceptual model.

Galya pointed out that the main goal is to replace some ICT survey questions and the idea is to go beyond the ESSnet. Galya suggested sending the conceptual model to Monica for checking and verifying and it would be better using the same tools and Albrecht recommended staying in the same terminology as WPF. Galya noted that the explanation of the conceptual model is envisaged in the chapter Big data processing life cycle on online-base enterprise characteristics – first subchapter: Conceptual production processes model (on the base of the BREAL Big data lifecycle). This is for WPC business layer. WPC team will prepare something specific for the package and send it to Monica, only this diagram for verifying. Albrecht recommended preparing of a very first approach and discussing it with Monica. Galya stressed that it is only a business layer, context, without IT layer.

Galya clarified that as regards target population for the URLs Inventory it is the same as in ICT survey. But for use-case 2 it should be derived population – only enterprises having websites. Aidan didn't agree. Even the statistics for businesses without websites is important. But Galya commented WPC countries try to make Big data official statistics. Galya stated that population for Big data statistics is different, it's selective population, not all population as e.g. census population. Big data expert defines target population. ICT variables definition was commented. It was agreed that modeling and interpretation should be linked to calculation of statistical indicators. When URLs Inventory is produced it follows going back to SBR and this is the link between use-case 1 and use-case 2.

There was a discussion about the timing of the cross-cutting WPs with respect to the output that has to be delivered by implementation WPs such as WPC. With respect to WPF is was noted that there is no mature document available yet. It was agreed that as long as the other work packages have nothing written down in a deliverable, the work of WPC on the reference methodological framework, at least version 1, cannot take the output of these work packages into consideration. Galya mentioned that we already lost 3 months and it was decided to start working on the document taking the output of the previous ESSnet (for example WP8) as a starting point. The leading partner for each chapter will put together a first draft, not exceeding 5 pages to keep it concise and clear, thus taking into account earlier comments of the review board on the content and length of such deliverables.

5. Wrap-up and planning next steps

Conclusions:

- It was agreed the text of the Methodological Framework document to be concise and clear;
- It is necessary to start writing the Methodological Framework document – there are 3 months left to have v.1 of it. The document is starting writing by leading country for each chapter, the others adding and a group of countries integrate chapter by chapter;
- WPC first deadline is for the Review board on 24th September 2019. Therefore **on 1st September 2019** the partners should have first draft version of Methodological Framework by chapters;
- The presentations from F2F meeting will be put on wiki and they will be the starting point for preparation of the Methodological Framework document;
- It was agreed to remove the chapter for legal aspects from Methodological Framework;
- Chapter 2: Business context is leading by AT;
- IE will contribute to the introduction and Chapter 3 - GBSPM and BD life cycle;

- Chapter 4: Reference architecture BG is leading because Kosta is part of WPF. As regards Chapter 4, Galya prefers to wait **till the end of July 2019** to have output from WPF and for the moment not to work on that chapter;
- Chapter 5: Now only 1st subchapter will be elaborated. 2nd and 3rd subchapter are postponed (implementation requirements and the set of recommendations) remains for Methodological Framework v.2.
- Chapter 6: Linking and integrating webscraped data – ISTAT will elaborate mainly. Donato will ask Monica about finding the right methodology.
- Chapter 7: Data quality and metadata will be done by PL (like a leading country) and AT.
- @Aidan (IE) will send a standardized template for experimental statistics of each country to be produced as a deliverable and will look at the methodological notes from the previous project on the wiki.
- @Galya invited Luke and Phil (UK) if willing to contribute to the introduction, business context and conclusion.
- All commented that on having real results of the work WPC has a real chance to become part of the Web Intelligence Hub of Eurostat.
- @Luke and Phil (UK) will contribute to the final proofreading of the document;
- @Galya will update the structure of the Methodological document and send it to the all WPC participants;
- All agreed not having a web-ex meeting in July but in August – around 20th, additional communicating of the date;
- For the next web-ex meeting in August there should be preliminary chapters.