WPJ WebEx Meeting

Data and time: 1 February 2019, 12:30 – 13:35 CET

Participants:

PL	Marek Cierpiał-Wolan, Dorota Jasiukiewicz, Piotr Szlachta, Teresa Matuła, Jarosław Napora, Mirosław Koszela
NL	Shirley Ortega-Azurduy + 1 person
IT	Mascia di Torrice
SK	Peter Mizik + 1 person
PT	Rui Alves
BG	Galya Stateva, Valentin Chavdarov
EL	absent
DE	absent

1. Opening and agenda

Marek Cierpiał-Wolan welcomed everyone to the second WebEx meeting of WPJ - "Innovative Tourism Statistics" package and briefly spoke about the agenda.
The main issues raised at the meeting were the methodology of web scraping and an overview of data sources delivered up to now by each WPJ partner. During the presentation the Polish team presented a PowerPoint presentation and an Excel file.
Representatives of Germany (Hesse) and Greece could not attend the meeting due to technical problems.

2. Methodology of web scraping

Piotr Szlachta (programmer) from Poland (PL) started the main part of the meeting. He gave a presentation on a proposal of web scraping methodology. He suggested dividing the web scraping process into 8 steps (i.e. websites identification, choosing the best websites – criteria, legal aspects, selection of adequate portals, implementation, cleaning and normalization, storage and data analysis). During the presentation, he explained each of these stages in detail. Piotr underlined the importance of cleaning and normalization of web scrapped data and that one should respect and follow all the rules set by a website while attempting to scrape it. He claimed that there might be a need to involve internal legal offices in some cases. The programmer

continued issues started on the previous WebEx meeting - in order to carry out the scraping activities, some software tools are needed to be developed and shared for using between the project partners. Now, it is known that the web scraping process is implemented in R language by the Netherlands and Germany.
In this context, Piotr also mentioned the differences in the data collected based on country of origin. He presented the experience of using the same script for hotels.com website by Polish team and Germany (Hesse) and results they obtained, respectively. For this case further investigation is needed to find out why the results differ. It will also help us to decide which method is better.
In the final part of the presentation, Piotr pointed out other problems encountered during the web scraping process currently performed by the Polish team, including structural changes in scraped portals that affect the continuity of the process, scraping of dynamic content websites and setting up appropriate delays and time to scrap data to avoid IP blocking.

After Piotr's presentation, the leader of package (PL) asked partners for their experiences based on previous work with web scraping. He also wanted to know the comments and opinions on using the script for web scraping sent earlier by Piotr to each partner country.
During the discussion, Mascia di Torrice (IT) said she talked with her colleagues from IT staff about the method they used in their office and passed Piotr's script to them. They did the web scraping process using HtmlUnit and JSoup. The following week they were supposed to look at the received script for hotels.com and present the results of tests using it.
Next speaker was Galya Stateva (BG), who wanted to know the name of the person responsible for web scraping from the Polish side, because they are trying to learn received script and they need some help in it. She pointed out the lack of experience concerning Big data sources for tourism statistics, however the Bulgarian team has a good experience in web scraping activities regarding enterprise characteristics during the previous ESSnet on BD I project, WP2. (at present, the Bulgarian team is working on webscraping on enterprise characteristics (WPC) and OJV (WPB)).
Representatives from Slovakia (SK), Peter Mizik, confirmed they tried to use the script prepared by Piotr. It worked, but the person responsible for this topic from Slovakian side was in the hospital at that time. Therefore, they will send feedback when he returns to work.
At the end of discussion related to web scraping, Marek Cierpiał-Wolan (PL) emphasized the need for increased work on web scraping methodology due to the necessity to prepare the first deliverable by the end of April, as it was scheduled in the project. It is a report entitled "ESSnet Methods for webscraping, data processing and analyses". As it was mentioned at WebEx meeting for WP leader on 23 January 2019, all workpackages should remember it is necessary that the Review Board is in the loop of every deliverable before deadline.

3. Overview data sources

This topic was presented by Dorota Jasiukiewicz (specialist in the field of tourism statistics) from Poland, who was showing a detailed table for gathered data sources prepared in an Excel file. At the beginning, she thanked everyone for submitting updated data according to suggestions from the last WebEx meeting (4 January 2019).

Dorota described the process of analyzing the list of data sources related to tourism statistics which had been already carried out. At the first stage, all of the received information on data sources sent by individual WP partner was just put in one sheet (named *Table (per country)* in the presented file). Then the Polish team tried to sort them out (using colours) into thematic areas according to data they include. Some of sources were not coloured yet, because of too little information about them. Dorota also mentioned that the presented catalog is quite a rough grouping and just the first step in order to specify later which sources can be used for estimations concerning supply or demand side. It is very important, because the goals of this workpackage are inter alia to obtain flash estimates of tourist and overnights spent as well as improve and verify estimation of tourist expenditures and the volume of tourist traffic. Moreover, she informed that apart from administrative data sources most partners sent also information on the statistical data sources concerning tourism. She noted that this information would be also very useful because the idea of Tourism Information System that might be developed during this project is integrating various big data sources with administrative registers and statistical databases. Dorota pointed out that the level of data availability is very diverse between countries and partners of WPJ should decide on the most effective approach, the best solution in this subject - the same level will be examined in all partner countries or different in each country. Other issues discussed by Dorota were: if administrative sources provide us with information about tourists and same day visitors and how to treat data sources which are marked as "not accessible" (will it be deleted as useless for the project or will it be treated as a potential source that will be available in the future?).

At the end of this part of presentation, the third table was shown in which Polish team gathered information about each type of source, taking into account the frequency and information provided, which can help us in direct comparison between all partner countries in this regard.

The grouped data sources catalog, presented by Dorota, including information from all countries will be sent to all partners with a request for comment, feedback and supplementing the missing information (e.g. gaps regarding to frequency, level, accessibility, etc.).

Discussion about work done on data sources sent by the partner countries was started by Marek Cierpiał-Wolan (PL). He hoped the analysis of the collected information would be improved in the

near future and the work related to combining them with data obtained from the web scraping process would be started.

Representatives from the Netherlands (NL), Shirley Ortega-Azurduy, spoke first and the clustering which had been done was a good idea, in her opinion. She asked for sending presented tables for a detailed look.

Rui Alves from the Portuguese team (PT) talked about data from credit cards. Next he stated that all information, including variable names and data sources, should be presented in English or should have bi-lingual version (English and native language). Another remark concerned the column "short/middle/long term source" and adopting format that could be more specific (for example: "Data available from [year]"). Future availability of sources should also be taken into consideration and information such as data ownership and publishing policy could be added.

Mascia di Torrice (IT) also referred to credit card data. In Italy, they are planning to organize a meeting with the Bank of Italy in order to obtain data on card payments, including the nationality of the person making the transaction.

Peter Mizik (SK) started by expressing appreciation for the work done on data source catalog. He also said that in the case of mobile phone data, they were negotiating with the administrators for data on this project. No detailed information on this issue yet.

During the discussion Galya Stateva (BG) informed the group she had already came back from Wiesbaden where she had participated in workshop "Trusted Smart Statistics: policymaking in the age of the IoT" . Answering question received from the leader of the project (PL) about detailed information, she pointed all workshop's materials are available at: https://ec.europa.eu/eurostat/cros/content/workshop-trusted-smart-statistics-policymaking-age-iot_en .

The representatives of the Polish side (Marek Cierpiał-Wolan) also answered the question asked by the partner from Bulgaria. Valentin Chavdarov (BG) wanted to know how the web scraping would improve the present statistics and how the data obtained from web scraping would be linked with existing statistics. WP leader (PL) confirmed that it was a difficult process, but his employees from the Mathematical Statistics Division were in the middle of the process of developing methods of combining these data and it would be presented and discussed next time.

4. Any remaining issues

There were no remaining matters and the meeting was closed. WP leader (PL) thanked everyone for the meeting and invited to the next WebEx which is set to 1 of March at 12:30 (CET)*.

After the WebEx meeting the list with contacts to persons taking part in the grant from each partner country, as well as the ones responsible for task 1 and task 2 and the presentation about web scraping methodology presented by Piotr during the WebEx meeting were sent.

* connecting from 12:15.