WP2 Face to Face meeting, Gdańsk, October 5-6, 2017

Participants to the meeting:

Jacek Maślankowski (GUS, Poland)
Matthew Greenaway (ONS, United Kingdom)
Olav ten Bosch (CBS, Netherlands)
Dan Wu (SCB, Sweden)
Giulio Barcaroli, Monica Scannapieco (ISTAT, Italy)
Maia Papazova (BNSI, Bulgaria)

The agenda of the meeting is:

Thursday, 5th of October, 2017

1. Welcome / Introduction / Overview 9:00
2. Country updates – WP2 related activities and software demonstrations (format 15 minutes presentation followed by 10 minutes discussion) 9:30
3. Coffee break 11:00
4. Discussion on deliverables for experimental statistics (tables, indicators – who can deliver the data) 11:30
5. Review of SGA-2 plans 12:30
6. Lunch break 13:00
7. Discussion on SGA-2 activities and deliverables 14:00
8. Close 17:00

Friday, 6th of October, 2017

1. Discussion on indicators and tables 9:00
2. Coffee break 10:30
3. Discussion on countries responsibilities for SGA-2 pilots 11:00
4. AOB 12:00
5. Lunch break 12:30
6. Close 13:00

Day 1 October 5th

After a short introduction by Monica Scannapieco, the following presentations were shown:

- Jacek Maślankowski presented the software written in Python for the social media pilot, focusing on the work done on it to be re-used by other countries. In addition he discussed the URL retrieval use case implemented by using the Istat's software. On the Job offers web scraping, he illustrated the interest by GUS to focus on administrations that have specific pages for job offerings.
- Maia Papazova presented for BNSI the work done within SGA1. First, she presented the URLs retrieval pilot. She mentioned the message published on BNSI web sites to inform enterprises about scraping activities (**Action 1**: Maia will send the link around). She also presented the results for the e-commerce use case. Then she illustrated the plans for SGA2, namely:
  1) Improvement URLs inventory
  2) Ecommerce testing Istat's software
  3) Social media improvement
  4) Candidature job offerings
- Giulio Barcaroli presented for Istat some initial work on the quality evaluation of the work done within SGA1 by focusing on (i) impact of response errors and (ii) the representativeness problem. Some considerations on the quality of aggregated figures were done and in particular on:
  - Comparison of the variance (sampling variance for survey and estimation of variance for the model)

- Simulation study: bias, variance, MSE. Adj. model-based seems to be the best.

A discussion followed on the need to provide clear methods for the evaluation of quality of aggregates, and on the need to take into account label noise in the machine learning step. Maia observed that the correct short name for the ecommerce use case is indeed "Sales facilities on Web sites" (**Action 2**: call the Ecommerce use case accordingly from now on)

- Matthew Greenaway started his presentation by giving good news: on the basis of the netiquette and the work done also by other countries within SGA1, there is eventually a positive decision by ONS to perform massive scraping. Matthew showed the results obtained so far on URLs retrieval, Ecommerce and Job advertisements pilots. The scaling up to massive scraping is envisioned for SGA2 for URLs retrieval and Ecommerce
- Dan Wu presented the recent acquisition by SCB of structured data on companies in Sweden by the VAINU company. A discussion followed on how to assess the quality of the provided data. A similar situation was experienced by CBS with companies DataProvider and TextKernel. Olav ten Bosch observed that such companies may have business models that are not useful for statistics.
- Olav ten Bosch started his presentation by citing (i) the recent US sentence that prohibits LinkedIn to block scrapers on public profiles. He underlined the issue on who owns the community data: the community (LinkedIn users) or the community provider (Microsoft) and (ii) the double face of ethical behavior in scraping (Scraper side and Site owner side). Then he presented the results achieved so far on the URL retrieval use case (peculiarity: Use of snippets), Ecommerce 8request to put it in production with runs twice a year). He cited a master thesis on classifying business activity with business websites using text-mining that could be used for the SGA2 use case on classifying businesses on the basis of NACE. Some plans for SGA2 include:
  - Scale up URLRetrieval to the Business Register
  - Ecommerce advancements
  - Social media presence

Point 4 of the agenda on Discussion on deliverables for experimental statistics was introduced by Monica Scannapieco and a discussion followed with the following decisions:

- For most indicators, the Eurostat Population for ICT survey: must be used for comparability. Link: http://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm
- The following indicators will be considered:
  - URL Retrieval - Rate(s) of retrieved URLs from an enterprises' list
    - IT;NL,BG;UK
  - Web sales - Rate(s) of enterprises engaged in websales from enterprises websites
    - IT;NL,BG;UK
  - Job advertisements - Rate(s) of enterprises that have job advertisements on their websites
    - IT,BG, PL
    - **Action 3:** Jacek will provide which is the reference population for this indicator computed by Poland.
  - Social media presence:
    - Rate(s) of enterprises that are present on social media from their websites
    - Percentage of enterprises using Twitter for a specific purpose, i.e.
    - a) Develop the enterprise's image or market products (e.g. advertising or launching products, etc);
    - b) Obtain or respond to customer opinions, reviews, questions;
    - c) Involve customers in development or innovation of goods or services

Gewijzigde veldcode

- d) Collaborate with business partners (e.g. suppliers, etc.) or other organisations (e.g. public authorities, non governmental organisations, etc.)
- e) Recruit employees
- f) Exchange views, opinions or knowledge within the enterprise
- UK;PL;IT;NL, BG . **Action 4**: Maia will check the availability of Bulgaria.

Point 5 of the agenda started as to discuss SGA2 activities and responsibilities.

Monica Scannapieco made a presentation on activities planned for SGA2. After a discussion, the following decisions were taken for the use cases already started in SGA1:

- Use case 1: URLRetrieval IT NL BG PL
- Use case 2: Ecommerce IT NL BG UK PL SE
- Use case 3: Social media IT BG PL UK NL SE
- Use case 4: Job advertisements IT BG PL SE

With respect to responsibilities assigned in SGA1, the following differences emerge:

- NL, UK will do UC3
- PL will do UC2
- UK will not anymore do UC4

With reference to the two new use cases the following decisions were taken

- Use case 5 SDG:
  - feasibility study (all)
- Use case 6 NACE:
  - feasibility study (all)

Two use cases are in the SGA2 workplan for being evaluated, namely: (i) web accessibility and (ii) support to ESBR. **Action5:** Monica will write to Eurostat for gathering details on that.

In terms of Task 4 – Future Perspectives: Testing information extraction techniques and Applicability of Findings, the following decisions were taken:

- Evaluation of extension of NLP techniques. **Action6**: Giulio will organize a meeting with Istat's people working on that.
- Evaluation of word embeddings. **Action 7**: Monica will organize a meeting with Istat's people working on that.

There was a discussion on next physical meeting for SGA2. Monica notified that on the basis of a request of Nigel Swier, coordinating WP1, it was decided that the meeting will not be a joint meeting between WP1 and WP2. After a short discussion, it was decided to have the meeting in Rome, March 15 and 16, 2018.

Day 2 October 6th

The second day started with a discussion of the workplan that was finalized as follows.

Tasks	October	November	December	January	February	March	April	May	Milestones
Task 1: URLs Retrieval			M1		M2				M1=First results of use case 1 (end Dec) M2= Final results for use case 1 + meth note (end Feb)
Task 2&3: Access and Analysis									
Ecommerce\Web sales						M3			M3= Final results + meth note (end March)
Social Media	M4		M5				M6		M4= Sharing sw by Jacek (mid Oct); M5= Test of Jacek sw (end Dec); M6=Final results (mid April) + meth note
Job vacancies						M7			M7= Final results + meth note (end March)
SDG		M8							M8= Sw sharing by Matthew (end November)
NACE Activities		M9							M9= Meeting (mid November)
Task 4: New Methods		M10							M10= Two meetings (mid November)
Deliverable writing down							M11	M12	M11= Deliverable version to send to reviewing committee (Mid april) M12= Final deliverable version (end May)
Analyses by Stat Sweden					M13				M13= Analyses by Stat Sweden on DB by VAINU

Action 8: Template on methodology to be prepared by Olav.

The dates of the two next virtual meeting were fixed on November 9Th 2017 and December 14th 2017.

The November 9th meeting will be about NACE usecase plans (M9).

The December 14th meeting will discuss results of Jacek sw tests (M1).

The final topic discussed was about Big Data Pilots II and how continuing the work done in Big Data Pilots I (SGA1+SGA2).

It emerged:

- Need for consolidation of approaches developed in terms of: (i) use for registers (quality+ enrichment) and (ii) use for surveys (quality+replacing variables+ variable enrichment).
- NACE, SGD, relationships with Euro Group Register are possible candidates for subsequent works.