

Minutes of the 4° Wp2 meeting – 1 September 2016, virtual meeting

Participants to the Meeting :

- ONS: Elizabeth Metcalfe
- CBS: Olav ten Bosch
- BNSI: Galia Stateva, Magdalena Miteva
- SCB: Ingegerd Jansson, Dan Wu
- GUS: Jacek Maślankowski
- Istat: Stefano De Francisci, Monica Scannapieco, Donato Summa

The following Agenda was scheduled for the meeting :

- 1) Presentation and discussion on the next steps:
 - Sharing URLs retrieval software procedure by Istat
 - Starting of scraping activities using the sandbox
- 2) Status of the report on legal aspects
- 3) Finalization of WP2 for SGA2
- 4) Preparation of November WP1/WP2 meeting
- 5) Any other business

For each Agenda item the following major points were discussed and related decisions were taken :

1) Presentation and discussion on the next steps

Monica said that the Italian software for URL retrieval will be shared among other participants with a readme file and a paper describing the work within which that software was used.

Galia said that they developed another SW for URL retrieval but will use also the Italian one in order to compare the performances between the two. A document about the Bulgarian software for URL retrieval was already uploaded on the wiki.

It was decided that the sharing of the developed software will be made on the project wiki but also on the sandbox environment and on GitHub.

There was a discussion about the Use cases implementation proposal and each Country applied for one or more use cases, the final situation is the following:

- E-commerce IT/UK/BG/NL
- Job advertisement IT/UK/PL/SE
- Social media presence UK/PL/BG

The implementation will start in November and the first results are expected by February.

2) Status of the report on legal aspects

There was a discussion about the progress that each Country made, briefly :

UK: Elizabeth reported that they have some problems mainly related to an “ethical” approach to access web sites, i.e. they will consider the case of not accessing web sites if the owners do not want to.

on the more appropriate way to access the sites

IT: Monica said that in order to scrape enterprises website it is necessary to declare this intention, for this reason Istat will put on its website an information note and this is enough from a legal point of view

BG: Galia said that BNSI already published a notice on its website so now they can scrape enterprises websites without legal problems

SE: Ingegerd said that they began to write something but essentially they are talking with a legal consultant before going forward

NL: Olav will update the document using the content of a presentation that he made on the topic

PL: Jacek said that a draft version of their part of the document will be available by the end of October, it will be finalized in November.

3) Finalization of WP2 for SGA2

Monica explained the criteria used for SGA2 cost repartition, essentially each Country except Italy received a flat percentage. The Italy percentage is a little bit higher because of the coordination task. All participants agreed.

During SGA2 other techniques beside the “bag of words” approach to the text information extraction task will be investigated , some cited examples could use ontology or deep learning techniques.

4) Preparation of November WP1/WP2 meeting

The November WP1/WP2 joint meeting will be held in Rome on 7/8/9. During the meeting will be shown the results obtained and there will be the demos of the URL retrieving software.

In particular on 7 November it will be held the WP1 meeting, on 8 November it will be held the joint meeting between WP1 and WP2 and on 9 November it will be held the WP2 meeting. The other participants agreed this scheduling because it facilitates the arrangement task for the participants of both WPs.

5) Any other business

Participants agreed to schedule the next WP2 virtual meeting on 6 October.

To Do activities :

- Istat will send the minutes of the meeting
- Istat will share the UrlSearcher software and a paper about URL retrieval procedure
- Each country will work on the report on legal aspect in order to have a first draft by end of October 2016.