THIS IS STILL A DRAFT VERSION TO BE AGREED UPON BY ALL PARTNERS

I Internal Meeting of WP5 Mobile Phone Data

Madrid, 7-8 June

Wed, 7 June			
09:00-09:30	Welcome and agenda adoption		
09:30-11:00	Description of compiled mobile phone data sets	FR, UK, DE, IT, NL, BE, FI	10 min. each + 20 min. debate
11:00-11:30	Coffee break		
11:30-13:00	Description of target statistical outputs	FR, UK, DE, IT, NL, BE, FI	10 min. each + 20 min. debate
13:00-14:00	Lunch		
14:00-15:00	Methodological proposals I: general framework, ecological sampling, and geostatistics	ES	
15:00-16:00	Methodological proposals II: admin data methodology and Heckman correction technique	RO	
16:00-16:30	Coffee break		
16:30-18:00	Debate on input data sets, outputs, and methodology	All	

Thu, 8 June			
09:00-11:00	Description of national ongoing/intended data processing	FR, UK, DE, IT, NL, BE, FI	15 min. each + 15 min. debate
11:00-11:30	Coffee break		
11:30-13:00	Internal technical reports: overview and questions Positium Sharemind HI – encrypted mobile Big Data secure processing	Positium	
13:00-14:00	Lunch		
14:00-15:00	Agreements and plan of action	All	

General Information

This draft agenda is structured upon the approach of (i) having a detailed description of the input mobile phone data sets we have been able to compile in the SGA-1, (ii) having a detailed description of the target statistical outputs we want to produce, and (iii) building a methodological driving us from the input data sets to the statistical outputs.

Regarding the contents of each item of the agenda, we mention:

1. Description of compiled mobile phone data sets
   We need to reach a homogeneous description of mobile phone data sets in order to single out the most important features for their later processing and methodological analysis. Each partner having access (or about to) will provide a short and concise description of their mobile phone data sets (10 min each). To achieve homogeneity and comparability, we request the following elements to be clearly spelled out in their presentations:
   - Format: are they relational databases, files, other?
   - Schema: in particular, the number of variables included in the data sets and the strict definition of these variables must be contained. This second requirement is especially important for some methodological aspects: how are identity, spatial and time attributes assigned to each record. Do they come from CDRs or signalling data are also used to compile the data set?
   - Volume and coverage: both the size in Gb (or Tb or Mb) and the geographical, time, and population coverage must be also included. Are data preprocessed so that particular registers are dropped out (e.g. machine-to-machine interactions)?
   - Software tools: is there any specific software tool especially required to transmit, store and process (if already processing) these data sets?

   Time for a short debate after all presentations has been allocated, although longer time slots are planned later on.

2. Description of target statistical outputs
   We need to reach a homogeneous description of the statistical outputs we want to produce, at least from a methodological point of view. So far, to our best knowledge, as included in the ESSnet application form, we target population size in tourism and mobility domains. In passing, daytime population size stands as a first compulsory intermediate result. Thus, we are targeting count statistics. We need to describe exactly what aggregates/indices we want to estimate and their level of breakdown, especially geographical and time breakdown.

   Time for a short debate after all presentations has been allocated, although longer time slots are planned later on.

3. Methodological proposals I and II
   As the skeleton of a generic proposal both Spain and Romania will present diverse suggestions to produce the target outputs. This proposal is detailed in one of the internal documents in preparation to be disseminated before the meeting.
4. Debate on input data sets, outputs, and methodology
   We include a time slot to debate around the proposed general framework going from data sets to target outputs.
5. Description of national ongoing/intended data processing
   In many cases, some processing work is on-going or already planned, potentially with some methodology in mind. In order to include all these initiatives in the generic proposal, we allocate time slots (15 min) for each partner having (or about to having) data so that they share their plans.
6. Internal technical reports: overview and questions. Positium Sharemind HI
   The third source to compile a methodological proposal is the report requested to Positium about data processing. Positium has been invited to the meeting so that firstly they will present an overview of the reports and answer questions related to these reports. Secondly they will present a novel technological solution to access and process Big Data in a secure way.
7. Agreements and plan of action
   The final item is devoted to reach a common agreement on the immediate plan of action to implement the agreed methodological proposals.

As a general comment we want to mention that the use of machine learning and data mining techniques going from microdata (at mobile phone level) to aggregated data (per municipality, per LA, etc.) is not currently well covered in the proposal under construction and thus in this agenda. If you feel you can make a contribution in this direction, please let us know to modify somehow both the proposal and the agenda.