# ESSnet Big Data

## Specific Grant Agreement No 1 (SGA-1)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
http://www.cros-portal.eu/.....

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2015.007-2016.085**

## Work Package 5

## Mobile Phone Data

## Deliverable 1.2

# Guidelines for the access to mobile phone data within the ESS

**Version 2017-06-20**


**Prepared by: David Salgado (INE, Spain)**

Ciprian Alexandru, Bogdan Oancea (INSSE, Romania)
Marc Debusschere (Statistics Belgium, Belgium)
Françoise Dupont (INSEE, France)
Pasi Piela, Ossi Nurmi (Tilastokeskus, Finland)
Roberta Radini (ISTAT, Italy)
Susan Williams (ONS, UK)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775

# Contents

1

# Introduction

The present document is the second and final deliverable corresponding to the SGA-1 of the work package on mobile phone data of the present ESSnet on Big Data. This deliverable aims at providing general guidelines for the ESS partners when conducting their own negotiations to have access to mobile phone data.

Negotiations with MNOs to access data generated in their frantic daily activity are remarkably complex for many reasons. One of our first global goals in this document is to convey this fact sharing those factors feeding this complexity. In second place, when considering negotiations with an arbitrary MNO at the whole ESS level, the situation is indeed more intricate because the actual circumstances for each operator and each country are clearly different.

As a general motivating comment, let us briefly digress on the definition of Big Data for Official Statistics, especially as far as mobile phone data are concerned. This digression is clearly rooted in the issue of institutional access to this data source for official statistics production. The widespread definition of Big Data by means of the 3Vs (Laney, 2001) later on complemented with more Vs (Normandeau, 2013) is well known and extensively used in any reference to Big Data. As a first reflection, mobile phone data share the characteristics of volume and velocity of generation, but on the contrary they are highly structured and no much variety is observed. However, in our view, in the realm of official statistics production there exist more important characteristics, especially in comparison to survey and administrative data.

In the data ecosystem of official statistics production we propose to parameterise the data sources in terms of the following features:

1. Data do not contain information of the (possibly sampled) data provider but of third people.

2. Data play a central role in the business of the data provider.

3. Data have at least one of the three Vs of the classical definition of Big Data.

Feature 1 reflects the fact that Big Data sources are indeed data about third people and they do not contain information about data providers themselves (usually large organizations, mainly private corporations). This is the case for mobile phone data, where information essentially regards mobile phone subscribers, not the operators. Feature 2 is a direct consequence of the digitalisation of the economic activity by which data is progressively playing a central role in the business processes. Thus they are key in the business strategies of economic agents. Sharing them has become increasingly sensitive to their activity.

Using these features, survey data would be characterised as having none of them. For example, turnover and number of employees in a short-term business statistics may be a piece of sensitive information for a business unit, but (i) these are data of the sampled business units, (ii) these variables do not play an essential role in their business activity, and (iii) no V appears in this kind of statistics.

On the other hand, administrative registers satisfy features 1 and 2 with regard to the Public Administration activities, but they do not fulfil feature 3. Finally, mobile phone data, web scraped data, smart meter data, ... all share the three features.

As you will hopefully conclude from this document, features 1 and 2 introduce such a novel ingredient in official statistics production that data access becomes an intricate issue arising in many entangled factors. Indeed, in some countries they even stand as an obstacle for the wide spreading of administrative registers in official statistics production.

To conform a set of guidelines for the access to mobile phone data we have identified three complementary subgroups. Firstly, mobile phone data are a vivid example of organic data (Groves, 2011). They are created by the activity of subscribers, temporarily stored, potentially reproduced for further business analysis, and finally deleted. Furthermore, the complex technological infrastructure (the cellular network), despite being subjected to rapid changes and evolution, follows a common set of design and implementation principles which the data generation, transmission, and storage relies upon.

All this portrays two immediate consequences regarding the access to data. On the one hand, the very concept itself of *mobile phone data* is fallacious: mobile phone data for statistical exploitation do not exist. In a telecommunication network there exist multiple information systems so that data are highly distributed. It is clear that at some level

data sets can be possibly found so that some further processing for statistical use can be conducted. This is the case of so-called Call Detail Records (CDRs), which contain details of the communication activity of each subscriber mainly for billing purposes. But much more information can be extracted from these networks with potential use for a better statistical exploitation. Thus, when firstly requesting access to data when negotiating with an MNO, it should not come as a surprise to find as a first reaction the question "Access to what data?".

On the other hand, the network complexity (especially its degree of distribution), together with the volume and velocity of generation of data brings to the front the issue both of extraction costs and of professional skills. The situation is complex because it depends very sensitively on the specific configuration of the network and the requested data.

Thus, our first chapter will focus on these technical aspects giving rise to these different issues surely appearing in any negotiation. It is not strictly necessary to be an expert in this technology to conduct a negotiation to have access to mobile phone data. However, from our experience, having certain background is noticeably beneficial for a better understanding with MNOs.

As a second subgroup of factors, there exist many entangled business considerations. Business considerations go from data extraction costs over legal requirements to communication, confidentiality and privacy issues. These (and others) were systematically debated during a workshop held in Luxembourg in September, 2016 among official statistics producers and MNOs. Chapter 3 contains a detailed description of this debate full of views, judgments, and perceptions from the different stakeholders.

Finally, as the third subgroup of factors, we understand that our experience itself in contacting and negotiating with MNOs could be useful as a guidance for others' initiative in their own country. In chapter 4 we include useful advices arising from our experience in requesting access to mobile phone data in our countries.

There is no definitive procedure to conduct a successful negotiation. There exist many factors involved. This document exposes as many as we have been able to identify in our experience.

2

# Technical guidelines


**Executive summary**

This chapter contains a description of the principles which the technological structure of a mobile telecommunication network relies on. Although it is not strictly necessary to bear an expert knowledge in this technology to negotiate access to mobile phone data, it is remarkably beneficial to understand several factors in the underlying complexity.

Mobile phone data for statistical exploitation do not exist in a cellular network. They are organic data created, reproduced, stored, and deleted in a frantic business cycle providing a telecommunication service, not a statistical service. A clear specification of the requested data must thus be formulated and negotiated.

A mobile telecommunication network has a nested hierarchical structure so that at the top basic data mainly for billing purposes are compiled whereas at the bottom (where multiple information systems are geographically distributed across the national territory) a wealth of technical data exists.

The core set of variables for statistical exploitation embraces (i) (anonymous) identification variables of each mobile device, (ii) time attribute(s), and (iii) geolocation attribute(s). The creation of these variables depends on diverse factors operating in the network. Complementary variables can also be extracted. A description of all these variables are included in the next sections.

From a wider perspective, the network complexity entails two immediate issues. On the one hand, extraction costs must be carefully taken into account, especially when confronted with the firm international principle in official statistics production of not paying for data for this purpose. On the other hand, new professional skills are needed for the staff dealing with this task. Both aspects are briefly tackled in this chapter.

Finally, all these technical aspects are summarised in a table as a collection of issues to be dealt with the MNOs in a negotiation ranging from the premises where data are to be processed over data coverage to network technology.

Not only when conducting statistical analyses on mobile phone data but also when requesting access to them do we need to be aware of the technical infrastructure in which these data are generated. Mobile phone data in general and for statistical purposes in particular are not a closed set of data possibly stored in files or in databases assembled in the daily routine production in a mobile phone network. When approaching an MNO to request these data for official statistical purposes, more often than not we will face the question about exactly what data we are requesting.

The more or less technical hindsight to the technological framework behind the generation of mobile phone data included in this chapter aims to help statistical officers in their quest to have access to these data for official statistics purposes and to understand this kind of reactions from our interlocutors at MNOs.

We do not intend to provide a carefully detailed technical description of a mobile communication network (see e.g. Sauter (2006); Mishra (2010); Sauter (2014)) but only to abstract out specific technical contents and to focus on those aspects impinging on the access negotiations and later statistical exploitation of the data. We already provided an approximate idea of what mobile phone data are and how they are generated in section 2 of deliverable 1.1 (Salgado et al., 2016). Here we bring some more pertinent aspects to the surface in relation with the negotiations for the access to the data. Nonetheless, mobile telecommunication technology has been changing rapidly in the last decades and will foreseeably continue to do so. Thus it is meaningless to go very deeply into these details.

Having said that, the abstract structure of these networks, independently of the concrete technological details, is clearly pertinent to our present purposes. This structure has remained more or less stable in this time and it appears to remain so in the near future. Its description is the main object of the present chapter.

Following the general bottom-up approach of the whole ESSnet we have pursued basing our results on on-going experiences as much as possible. In this sense, since the production of statistics based upon mobile phone data is fairly new within the ESS, we have resorted to the Estonian company Positium, already producing this sort of statistics and involved in the preliminary work in the ESS related to mobile phone data for statistical purposes. In consonance, this chapter is heavily based on an internal technical report within the work package on mobile phone data prepared under specific request to Positium (Positium, 2016). Some parts of the forthcoming text has been adapted from this source to the present deliverable.

## 2.1 Overview

We must never forget that *the interest for mobile phone data for the production of official statistics does not reside in the mobile devices themselves but in their potential to measure and analyse human (and possibly business) populations*.

The link between humans as statistical units of analysis and mobile devices as objects of observation is firmly established in the current penetration figures of these handheld computing devices in European countries (Eurostat, 2016).

As we put forward in deliverable 1.1 of the present work package (Salgado et al., 2016), mobile devices can provide highly valuable information at three fundamental levels: (i) the (anonymous) identification of each mobile device, (ii) the geolocation of each mobile device, and (iii) the temporal location of each mobile device. Additionally, there may be some more complementary data.

These fundamental levels arise from the structure itself of mobile telecommunication networks. In the original wireless communications, the basic scheme was to establish a channel using one or two radio frequencies. The communication was possible by means of high-power radio transmitters connecting both parties. This entailed two major drawbacks for their extensive use in society: (i) each operator was legally limited to use a number of radio frequencies thus strongly impeding the extension to the current scale of users, since frequencies were reserved for each communication across the whole territory and (ii) high-power radio transmitters need powerful sources of energy (batteries) to cover the whole territory thus handicapping the portability.

In the 80s these problems were overcome with the current abstract structure of mobile networks. On the one hand, the territory under the service provision by the operators was divided into geographical cells with lower power transmitters and antennae so that interference of use between different cells was removed. Cells were connected by means of computer networking to allow subscribers in different cells to communicate. In this way, the limited number of radio frequencies of each operator could then be reused all across the whole territory thus paving the way for higher numbers of concurrent users.

On the other hand, lower power transmitters do not need such energetic demands and batteries can be much smaller (apart, of course, from the miniaturisation innovation itself of these power sources). This unchained the unprecedented levels of portability of mobile devices, currently limited basically by the size of keypads and touchscreens for the interaction between humans and devices.

These drove us to the current situation in which nearly everybody carries his/her

own handset (even two in some cases). Regarding the power supply, we observe no direct consequence regarding access to mobile phone data for statistical purposes. However, regarding the cellular structure of mobile networks, the fundamentals of this solution show a direct effect not only for having access to data but also for their later statistical exploitation.

The following sections go all the way from the mobile devices themselves (our object of observation) carried by humans (our object of analysis) to final potential data sets possibly accessed by NSIs to produce official statistics.

## 2.2 Mobile devices and SIM cards

Mobile devices are portable computing devices equipped with radio transmitters enabling them to connect to a telecommunication network (Sauter, 2014). Additionally, these devices are increasingly equipped with a number of sensors of different nature (accelerometer, gyroscope, digital compass, GPS radio transmitter, ...) (Khan et al., 2013). Depending on the amount of technology (these sensors) and their size, these devices receive diverse names (mobile phones, feature phones, smartphones, tablets, ...).

Apart from the radio transmitter, the connection to a telecommunication network is enabled through a so-called Subscriber Identity Module (SIM) card, which is essentially an internal integrated circuit card (ICC) providing diverse functionalities to establish an authenticated secure communication between the mobile device and the network (Sauter, 2014).

It is important to underline that this combination of mobile device and its SIM card is the origin of the diverse mobile phone data we are requesting access to.

To understand the link between subscribers (our objects of analysis) and the generation of their data, it is convenient to know that the SIM card provides authenticated identification in the network at two levels:

(i) To access reading, writing, and use of a SIM card a first security functionality enters into play. As an ICC (e.g. a credit card), a combination of evidence of identity stored in the card (e.g. a key) and personal information only possessed by the user (e.g. a password) will grant access to the SIM card. This is the usual PIN number introduction procedure to turn on a mobile device. Indeed, this is further protected by a second layer of security through a second password (PUK number) when the first procedure fails. When many PUK numbers failed to be correctly introduced, the card will be permanently blocked.

(ii) Once the SIM card is accessed, the diverse functionalities of the card are ready to use, especially the identification and authentication of the subscriber in the network.

To do this, the operator, when the subscription is formalized, programs the SIM card storing a so-called International Mobile Subscriber Identity (IMSI) code in it. This code identifies the **subscriber** across all networks (even possibly in other countries). It is a numeric code of length 15 at maximum formed by three parts: (i) the Mobile Country Code (MCC), which is a 3-digit code identifying the home country of the subscriber; (ii) the Mobile Network Code (MNC), which is 2 or 3-digit code identifying the national network; and (iii) the Mobile Subscriber Identification Number (MSIN)[1], which is a 9-digit code identifying the **subscriber**.

It is important to underline that the IMSI is not changed when the same SIM card is used in another mobile device. Thus a subscriber is (anonymously) identified in a network even despite he/she updates his/her mobile device.

It is also important to emphasize that a SIM card does not store the mobile phone number of a subscriber. Thus, when a subscriber changes the operator providing the mobile services and a new SIM card is programmed for the services in this new operator, the IMSI code also changes. However mobile phone number portability, a legal right for MNOs' subscribers when they change their operator in many countries, allows us to keep track of them through the so-called Mobile Subscriber Integrated Services Digital Network (MSISDN) code, i.e. the mobile phone number. There exist more identification codes associated to a subscriber related with different circumstances (roaming, moving from cell to cell, etc.; see [SECTION] for further details). It is important to keep in mind that this interplay of identification codes will allow us to anonymously identify s subscriber among different data sets for statistical purposes. Though apparently obvious, we must state that if a given mobile device with a given SIM card is used out of a sudden by another person, there will be no way to detect this change of statistical unit (apart from the data pattern analysis itself).

For completeness' sake, let us state that mobile phones, independently of the SIM card, have also a unique international identification number, i.e. the so-called International Mobile station Equipment Identity (IMEI) code and, complementary, the International Mobile station Equipment Identity Software Version (IMEISV) code. These

[1]When focusing on a single country, the MCC can be dropped out and the combination of the MNC and the MSIN is usually referred as the National Mobile Subscriber Identity (NMSI).

identities are not permanently related to a subscriber, since devices retain their IMEI and IMEISV even after transmission to another user either in the same network or to another network.

## 2.3 Diverse cellular technologies

From the inception and deployment of the first cellular technology in the 80s, the innovation has been non-stopping. Currently, MNOs operate using simultaneously diverse technologies from GSM[2] (2G) over UMTS[3] (3G) to LTE[4] (4G) (Mishra, 2010; Sauter, 2014), but the innovations go on and 5G is under way (Osseiran et al., 2016). They all have evolved from the same core cellular structure of GSM networks.

On the one hand, in this deliverable these diverse technologies will be abstracted out to focus on the GSM-like cellular structure of mobile networks. It is this structure which strongly influences how access to data is to be agreed upon between NSIs and MNOs. For access purposes, except for data volume issues (see immediately below), there is no deep implications arising from the differences in the technologies.

On the other hand, as stated above, different technologies bring successively higher volumes of data since the technical capabilities evolve producing more data when providing services to subscribers. However, being this a desirable feature for the statistical exploitation (we would have more data), this also brings more complexity in the data extraction stage (see 2.7).

## 2.4 Architecture of GSM-like networks

A GSM-like mobile telecommunication network is a collection of three nested types of subsystems:

a) The radio network or Base Station Subsystem (BSS), providing all elements to connect the mobile devices to the network over the radio interface (also known as air interface). It is here where the connection between mobile devices and antennae takes place.

b) The core network or Network Switching Subsystems (NSS), providing all elements for switching of calls, for subscriber management and mobility management. The core network may be optionally complemented with the Intelligent Network

[2]Global System for Mobile Communications.
[3]Universal Mobile Telecommunication System.
[4]Long Term Evolution.

(IN) subsystem, providing optional functionalities to the network (as the prepaid services, to name the most important).

c) The Network Management System (NMS), monitoring and managing diverse aspects of the network such as maintenance works (software upgrading, collection of statistics about perfomance, customer billing, ...).

This hierarchical nested structure is represented in figure 2.1. As a general theme, we will see that the closer to the mobile devices themselves data are generated, the more effortful data extraction will result (see section 2.7). It is thus important when negotiating access to mobile phone data to have some knowledge of where they are generated and stored.

Now we describe in more detail the basic functionalities of each subsystem with a special emphasis in those databases from which data for potential statistical exploitation can be extracted. Later on, we will detail which data of interest are generated and stored in each subsystem.

### 2.4.1 The radio network (BSS/NodeB/eNodeB)

A radio network is composed basically of (i) Base Transceiver Stations (BTS) and (ii) a Base Station Controller (BSC) (see first layer in figure 2.1 where two radio networks are depicted).

BTS is original GSM terminology. For UMTS and LTE technologies, their equivalent terms are NodeB and eNodeB. For our current purposes, we do not make distinctions. For future technologies such as 5G, the situation will presumably be similar. Indeed they all can also jointly referred to as Base Station (BS).

A BTS also comprises diverse elements, apart from the antenna itself, with diverse functionalities (encrypting and decrypting communications, spectrum filtering tools,...) (Sauter, 2014).

It is important to understand that each BTS is associated with a geographical cell covering the territory of the so-called Public Land Mobile Network (PLMN), i.e. of the entire network of the operator. Depending on the configuration of the BTS, these cells can be omnidirectional (sector angle of 360°), two-sector (each sector angle of 180°), three-sector (each sector angle of 120°) or four-sector (each sector angle of 90°). The so-called azimuth gives the orientation of each cell sector. See figure 2.1.

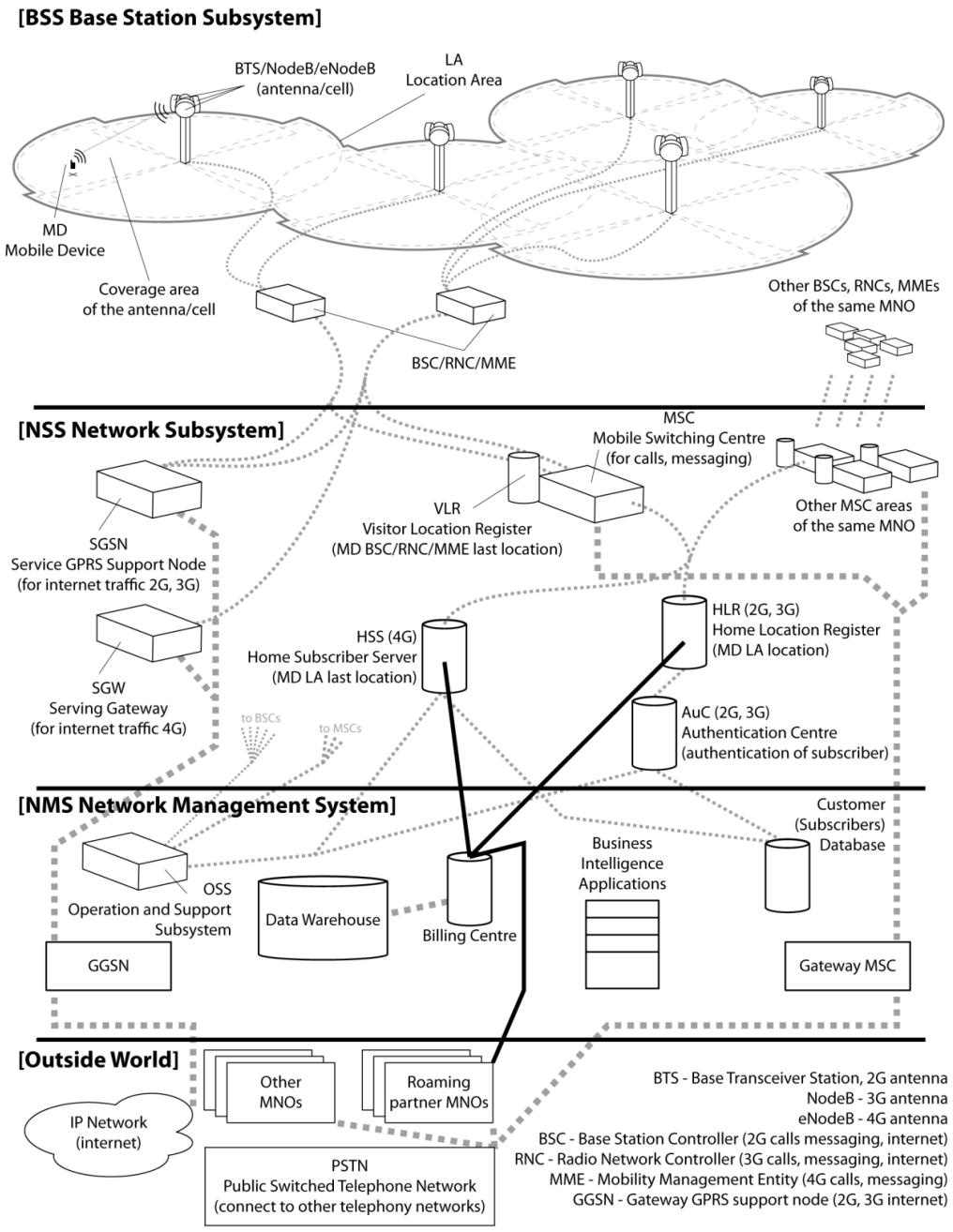


**Figure 2.1** Generic architecture of MNOs' information systems (taken from Eurostat et al. (2014))

A BTS, in principle, can cover a cell of radius up to 35 km. However, since a BTS can only serve simultaneously a limited number of mobile devices, in practice the cells vary due to a number of factors, especially the population density (but also e.g. the lanscape, the weather,...). Thus in urban areas they have typically a radius of 3-4 km for residential and business zones and in densely populated areas (shopping centers, downtown streets,...) they can be reduced down to 100 m. In rural areas, a priori less densely populated, the low transmission power sets a practical limit of 15 km.

Therefore, the configuration of cells is not a simple disjoint partition of the geographical territory and overlapping and intersecting cells are often found in practice. Furthermore, since power transmission differs for the different technologies (2G, 3G, and 4G), the size of the cell of a given BTS corresponding to each technology also differs and this must be taken into account when considering the geolocation attribute of mobile phone data (see section 2.6.3).

Several BTSs (of the order of tens or even hundreds) are connected to and managed by a Base Station Controller (BSC), which provides the control over each BTS as well as the connection to the core network.

Again, this is original GSM terminology, being Radio Network Controller (RNC) and Mobility Management Entity (MME) their equivalent terms in UMTS and LTE technologies, respectively (see figure 2.1). Their role in each case is similar.

Adjacent BTSs (thus adjacent cells) are grouped together into a so-called Location Area (LA) (see figure 2.1). This is intended to optimize the capacity of the network so that only mobile devices moving from one cell to another distant cell outside their LA communicate their updated position to the core network thus not producing overloading. Again this means that information about mobile devices moving from cell to cell inside the same LA will not reach the core network so it will be more effortful to access these data for statistical purposes.

### 2.4.2 The core network (NSS)

The core network takes responsibility for subscriber management, call establishment, call control and the routing of calls between different LAs and other networks.

The fundamental element of a NSS is the Mobile Switching Center (MSC), responsible for routing voice calls and SMS as well as mobility management in general and billing. It also acts as an interface between radio systems and fixed networks.

In a PLMN there are typically several MSCs. For our objective of access to data for statistical purposes MSCs are a vital element of the network, since it manages mobility receiving continuously subscriber locations (also referred to as location updates) to direct incoming connections (such as calls) to the right BSC, which in turn starts the process of paging[5] within the BSS.

For the communication between two different PLMNs (i.e. networks of two different operators), a gateway is needed to connect them. Such a gateway (SGW) is usually implemented in an MSC, which is then referred to as a G-MSC. For GPRS networks, the Serving GPRS Support Node (SGSN) handles the tasks listed here for the MSC.

Each MSC has a Visitor Location Register (VLR) associated to it. A VLR is a database of those subscribers having roamed into the LAs served by the corresponding MSC. Thus it is important to know that each subscriber in the network is exactly served by only one VLR at a time. As subscribers move all across the network, they are included and excluded from the corresponding VLRs. This makes VLRs a priori good targets to access data for statistical purposes.

The main information stored in VLRs is:

- IMSI codes.
- Authentication data.
- Services the subscriber is allowed to access.
- MSISDN, i.e. mobile phone numbers.
- Subscriber's HLR address.
- Access point (GPRS) subscribed.
- SCP Address (for prepaid subscriber).

The three first points are indeed a copy of the information stored in the SIM card of each subscriber.

MSISDN stands for Mobile Station International Subscriber Directory Number and it is indeed the subscriber's mobile phone number. It is composed of the country code, the national destination code and the subscriber number. It is important to remind that the MSISDN, due to portability of mobile phone numbers, remains unchanged when a

[5] Paging is the process of identifying the exact cell inside a LA in which a mobile device is located.

subscriber changes his/her operator thus also changing the SIM card (thus the IMSI). This association of IMSI and MSISDN also stands as an interesting feature for statistical purposes.

As stated above, VLRs get updated as subscribers change their location across the network. The information of each subscriber is taken from the Home Location Register (HLR). The HLR is a central database containing detailed information about each network subscriber who is authorized to access the core network. There can be several logical and physical HLRs per PLMN (i.e. per operator), but only one international IMSI/MSISDN pair can be associated with only one logical HLR at a time. This is also an interesting feature for the univocal (anonymous) identification of subscribers in the networks.

To close the description of an NSS, let us mention the Authentication Centre (AuC), which implements the function to authenticate each SIM card that attempts to connect to the core network (typically when the mobile device is powered on). Once authenticated, the HLR begins to manage the SIM and the corresponding services.

All preceding terms belong to GSM terminology. While not having changes in UMTS terminology, for LTE the NSS level brought new gateways as well as a counterpart to the HLR with similar functions:

- The Packet Data Network Gateway (PDN-GW) is used to give subscribers access to the Internet but can be also used to gain access to intranets.
- The Serving Gateway (S-GW) entity is used to manage connections between eNode-Bs and the PDN-GW.
- The Home Subscriber Server (HSS) is the counterpart of the HLR in LTE systems. In order to make roaming from one radio access network (e.g., UMTS) to another (LTE) more seamless, the HLR and HSS can be physically combined. Among other things, the HLR/HSS holds information about user IMSI, user's telephone number (MSISDN), ID of the current serving MSC, SGSN and MME.

### 2.4.3 The network management subsystem (NMS)

The purpose of the NMS is to monitor and to manage various aspects of the network. The functions of the NMS can be divided into three categories: fault management, configuration management and performance management. The NMS handles these tasks system-wide, starting from single BTSs up to MSCs and HLRs. During these tasks, the NMS also collects and stores data about occurring problems, configuration status

and the performance of individual network elements. These data are later used by the MNO to analyse the performance of the network.

- The Operation and Support Subsystem (OSS) takes responsibility for maintenance works such as software upgrading and the collection of statistics about network performance. The OSS also contains databases that hold information about network elements, such as locations of cell sites, i.e. of BTSs. This is a relevant feature possibly impinging on the geolocation dimension of the requested data to MNOs.
- The Gateway GPRS Support Node (GGSN) acts as an extension for the SGSN (see description for MSC) in GPRS networks to connect a GPRS network to an external data network (e.g., Internet).
- The Billing Centre (BC) is responsible for gathering network usage data for every subscriber. The data is pulled from the MSC (SGSN and GGSN are also involved in generating billing data).

## 2.5 The billing process

In order to be able to produce a bill for each subscriber, the MNO needs to maintain charging mechanisms that are responsible for producing and combining Call Detail Records (CDR) and Internet Protocol Detail Records (IPDR), which are subject to billing information generation (e.g. applying service rates). MNOs do not have one specific entity that is dedicated to the CDR or IPDR generation. Instead, every entity that is responsible for providing a certain service generates CDRs/IPDRs. As a matter of fact it is not rare for multiple network entities to be involved in CDR/IPDR generation processes related to one service type.

From entities described in previous sections, the HSS/HLR, SGSN, PDN-GW are examples of entities that create CDRs/IPDRs. Since CDRs/IPDRs are generated by different network elements, MNOs need a process that would be able to combine all this information. This process is referred to as mediation.

In addition to home subscriber records, CDR and IPDR files contain records about visiting (roaming) subscribers (inbound data). These CDRs and IPDR will be included in the Transferred Account Procedure (TAP) that will be sent to roaming subscribers' PLMN (outbound data).

Charging mechanisms can be generally divided into two broad categories: offline and online. Offline charging does not affect services provided to the users, while online charging constantly keeps track of the amount of services used and, when necessary, refuses further service for the user. For example, MNO monitors users with pre-paid

cards to determine whether they have used up their credit. During offline charging, chargeable events will be attached to a file with a size limit. Every time this limit will be achieved, a new empty file will be generated and the previous file becomes available to the mediation system, which will then forward this to the Billing Domain (BD) for further processing and analysis. On the other hand, online charging is much more complex compared to offline charging. In case of online charging, the account balance management function monitors the subscriber's account balance and grants permission for service based on that. During online charging, the CDR/IPDR generation is seen as an additional element, not requirement, which means that not all MNOs gather CDRs/IPDRs generated during online charging.

| Identifier | Description | Storage | Usage limitations |
|---|---|---|---|
| International Mobile Subscriber Identity (IMSI) | • Unique identifier for every mobile subscriber; not necessarily for every person/client.<br>• IMSI serves as the root of the subscriber data pseudo-tree within the same network (MNO). | • SIM card.<br>• HLR.<br>• VLR.<br>• SGSN.<br>• GGSN. | • Each new SIM implies a new IMSI (e.g. loss of SIM card, switching MNO,...). |
| Mobile Subscriber Integrated Services Digital Network Number (MSISDN) | • Mobile number of subscriber. | • HLR.<br>• VLR.<br>• SGSN. | • There can be 1:N relationship between the IMSI and MSISDN in the HLR.<br>• The MSISDN can be changed without changing the user's SIM card or the IMSI related to it. |
| Mobile Station Roaming Number (MSRN) | • Temporary identification number to route calls directed to a mobile device (within the BSS).<br>• Possibly identical to the MSISDN. | • VLR. | • Possibly more than one MSRN simultaneously per IMSI.<br>• Pool of MSRNs is used, i.e., at different times, different subscribers could have the same MSRN. |
| Temporary Mobile Subscriber Identity (TMSI/P-TMSI) | • Identity most commonly sent between the mobile and the network; randomly assigned by the VLR to every mobile in the area when switched on.<br>• Local to a LA, and so it has to be updated each time the mobile moves to a new geographical area. Thus together with location area identity, TMSI can identify subscriber uniquely. | • VLR.<br>• SGSN. | • One subscriber can have two distinct TMSIs – one for services provided through the MSC (TMSI), other for services provided through the SGSN (P-TMSI).<br>• It possibly changes every time a mobile device contacts the network.<br>• It changes every time a subscriber roams into a new location area.<br>• Pool of TMSIs is used, i.e., at different times, different subscribers could have the same TMSI. |
| Local Mobile Station Identity (LMSI) | • Temporarily identity possibly assigned to a mobile station visiting another network than its home network (e.g. when roaming in a foreign country).<br>• Allocated by the VLR if the location is updated and sent to the HLR together with the IMSI. However, the HLR does not make use of the LMSI. It includes it together with the IMSI in all messages sent to the VLR concerning that mobile station.<br>• Auxiliary searching key to speed up queries. | • VLR.<br>• HLR. | • Can be changed every time a mobile device contacts the network.<br>• Changes every time a subscriber roams into a new location area.<br>• Pool of LMSIs is used, i.e., at different times, different subscribers could have the same LMSI. |

**Table 2.1** Comparison of subscriber identifiers. Adapted from Positium (2016)].

Attributes needed for charging are another aspect of the matter. Since each service type has its distinctive features, CDRs/IPDRs also differ in view of the attributes which are necessary to produce a bill for a subscriber. For this reason, technical specifications exist for charging related to different service types. For example, call duration is an important attribute for an outgoing call but for sending an SMS attributes of this kind are redundant.

## 2.6 Mobile phone data

When negotiating access to data it is important to remind that mobile phone data, in the usual sense of traditional data collected and processed for official statistics production, do not exist per se. In the daily operation of a mobile telecommunication network the fundamental entity is the event, i.e. each action initiated by or targeted to a mobile device within the network. These events are the very beginning of a cascade of processing actions all across the network producing different kind of data, not all of them stored or easy to access.

Immediate example of these events are the use of services such as calling, sending text messages or connecting to the Internet. But there exist more like a change of cell or a change of LA because the mobile device is moving. The kind of data generated and stored for each event is specific to the type of event and their storage and processing may depend on the operational setting of the MNO.

For statistical purposes we will concentrate of the following minimal attributes associated to each event: (i) subscriber identifier, (ii) time attributes, and (iii) geolocation attributes. Needless to say, more attributes can (and should) be added to increase the value and potentiality of data for their statistical exploitation.

### 2.6.1 Subscriber identifier

As stated at the beginning of this chapter, it is human populations and not mobile devices themselves which are the target of the access, processing and analysis of mobile phone data. We are using mobile devices as a proxy to humans. Thus it is essential to understand how each human is identified not only within a given network but even across several networks. Nonetheless, there exists an increasing number of machine-to-machine communications which in the near future will certainly provide valuable information for statistical analysis. In this document, for the time being we shall focus on human behaviour.

During the time span of the analysis a subscriber identifier should be unique for each subscriber and as stable and permanent as possible. Complementarily, each person

should ideally have only one identifier (although people actually use several mobile devices thus the relation person:subscriber is indeed 1:n in some cases).

All across the network there exists a number of identifiers for each subscriber and/or mobile device. In some cases, the identifiers are local within the network entity processing the information (e.g. the BSC) and not stored or possibly reused to optimize operations. The two most important identifiers held constant across the network are the IMSI and the MSISDN (see section 2.6.1). In table 2.5 we include both temporary and permanent identifiers together with a description, the network entity where it is stored and potential usage limitations.

As important remarks:

(a) There exist multiple ways to univocally identify a subscriber in a network.

(b) Complementarily, the MSISDN (the phone number) and the portability of these numbers among MNOs can help us potentially identify a subscriber even among different networks (although some assumptions are made here to be contrasted with the data themselves).

To end about identifiers, the equipment identifiers such as IMEI and IMEISV show a priori little value for the statistical analysis of human population, since lately people often change their mobile devices buying new ones. This information, although considered temporary, might be stored in the HLR, SGSN or VLR.

### 2.6.2 Time attributes

Time attributes probably the more straightforward piece of data although with an essential value for statistical purposes. The international standard defines three different types of timestamps:

- Seizure time is the time when resources are seized to provide service to the subscriber. This field is mandatory only for calls that were unsuccessful.
- Answer time is the time when a call is answered – the connection was successful. Answer time is a mandatory field for successful calls.
- Release time is the time when seized resources are released again. Release time is an optional field.

Time attributes are generated by those network elements providing specific services. They are typically not stored in entities such as the HLR or VLR but are rather accessed during the generation process of data for billing purposes by using specific functions

triggered when a chargeable event occurs. For access negotiation and later affordable data extraction, it thus must be taken into account that not all events are assigned a timestamp possibly stored and recovered for later statistical purposes. At least not in standard operation conditions, thus a technological investment may be necessary.

It is important to pay attention to the time zone used in the timestamps. According to MNO's needs, default local MNO plus an offset may be under use, which may be possibly at odds with data received from other MNOs because of subscriber roaming in their territory. It must be discussed and checked in each case with the MNO the coherence of timestamps. As a recommendation from experts, time attributes should be presented in epoch[6].

### 2.6.3 Geolocation attributes

Geolocation attributes of mobile phone data are strongly dependent on the cellular structure of mobile networks detailed in section 2.4. By and large, we must make clear that in telecommunication standards geolocation identities are typically mandatory during the generation of the billing process. However, there exist circumstances in which this is not so, e.g. during international roaming outbound data might not include geolocation-related identities other than the country code.

Thus, as a generic rule of thumb, for domestic and inbound roaming data the geographical reference is a priori the cell identity together with the coverage area of each network cell. For outbound roaming data, the geographical reference a priori is the country of the roaming partner MNO. Occasionally, these outbound data may also include the network cell ID of the event. This must be discussed and agreed between the NSI and the MNO.

Each cell served by a BTS in a BSS (see section 2.4.1) can be univocally identified within a network by a so-called Cell Global Identity (CGI). This is a code composed of the so-called Location Area Identity (LAI) or Tracking Area Identity (TAI) (for LTE networks) and a Cell Identity. For packet-switched networks, Routing Area Identity (RAI) plays the same role as LAIs/TAIs. The LAI/TAI/RAI is in turn a compound code formed with the Mobile Country Code (MCC), the Mobile Network Code (MNC) and a Location/Tracking/Routing Area Code (LAC/TAC/RAC). The latter is a 2-octet long hexadecimal code univocally identifying each LA/TA/RA in a network. These unique CIs are commonly stored in a database at the OSS level (in the NMS).

[6]Number of seconds elapsed since 00:00:00 Coordinated Universal Time (UTC), Thursday, 1 January 1970.

Both LAI/TAI/RAI and LAC/TAC/RAC are temporarily stored in the VLR and SGSN. It is important to know that these registers (VLR and SGSN) are used and accessed in the billing process.

We see that geolocation attributes are strongly dependent on the cellular structure of the network, which in turn can change frequently (e.g. addition of new antennae, directional changes for individual cells,...). Thus it is important to agree with MNOs how these changes are to be included and managed in the process of accessing their data.

Finally, for statistical purposes these CGIs must be attached to geographical coverage areas. The minimum piece of information in this sense is the geographical coordinates of each network antenna. Depending on the accuracy of the statistical outputs, this information may be enough (e.g. working at country-level, NUTS3, LAU1). But if finer estimates are to be produced (LAU2, LAU3, municipalities, city districts,...), we also need the coverage area of each antenna. Two options arise: (i) either the MNO provide the geographical coverage are of each antenna or (ii) a theoretical cell coverage area computation must be undertaken. The former case is part of the negotiations with the MNO; the latter case, on the contrary, is part of the data pre-processing stage (to be dealt with in the next deliverable).

### 2.6.4 Additional attributes

Three main groups of additional attributes may be considered when requesting access to mobile phone data, namely, (i) event data, (ii) network data, and (iii) subscriber's additional attributes.

#### Event data additional attributes

Apart from subscriber identity, time and geolocation attributes, there exist more attributes to be potentially include in the mobile phone data sets. International specifications for charging divide them into four categories:

1. Mandatory attributes.

2. Conditional attributes depending on the fulfillment of certain conditions.

3. Operator-provisionable mandatory attributes which MNOs have provisioned to be always included.

4. Operator-provisionable conditional attributes which MNOs have provisioned to be always included if certain conditions are met.

Many of these attributes exhaustively included in the specifications show little value for statistical purposes, but a few of them are being commonly considered:

- Record type – helps to differentiate between data types (CDR and IPDR), service types (e.g., SMS, MMS, outgoing call, etc.) and distinguish between events that are not subscriber generated (e.g., location updates).
- Call duration or data volume for analysing mobile phone usage patterns.
- Subscriber or equipment identity for receiving/sending parties.

**Network data additional attributes**

The OSS usually stores additional information for maintenance and performance analysis purposes. This information can be highly valuable for improving the preceding attributes, in particular, attributes of the cells can be of great value.

**Subscriber additional attributes**

Databases with information of their customers are kept for diverse reasons. This information is highly specific to each MNO although typically includes:

- Socio-demographic characteristics of the subscriber (the contract holder) possibly comprising age, gender,... Notice that in some cases (corporative legal entities as contract holders) this information is apparently more difficult to attach to concrete people.
- Details about the contract and service: personal or corporative customer, invoice address, contract type (pre-paid, machine-to-machine, etc.).

Needless to say, this information is extremely sensitive regarding customers' privacy. The value of these attributes lies in providing extra dimensionality to the statistical analysis.

This information is not usually contained in the information systems of the network but in the Customer Relationship Management (CRM) system. These data must be used with caution in the statistical analysis since the connection to the concrete person using the mobile device is not immediate.

## 2.7 Data extraction

Data extraction from MNO's information systems depends on the specific technical solutions used by MNO for their production. In general, every MNO has a billing centre and, most likely, a data warehouse where data is periodically stored for billing and further analyses for network planning and management. Besides central storage systems where data are readily available (section 2.7.1, it is possible to extract data in the process of probing at the BSS and NSS levels 2.7.2.

### 2.7.1 Central Storage Systems

Data extraction from central storage systems is the easiest way to obtain data for national-level statistics. There are several places where such data can be found, collected for different purposes and containing different kinds of information:

- The billing domain (BD) stores CDRs and IPDRs for charging purposes. Data will be stored here after a successful charging record generation procedure.
- Customer databases contain information about users, which can add extra value to CDRs or IPDRs data – e.g., socio-demographics and place of residence (where applicable).
- Data warehouses host collections of data from different sources but might not always be easily accessible due to the huge amount of data stored there.

Databases in the billing domain are generally considered most easily accessible. Data stored in the BD is gathered by a so-called mediation system from different network entities responsible for providing various types of services.

### 2.7.2 Probing

There are two types of probing — active and passive -– for monitoring and acquiring data from mobile telecommunication networks. Active probing is used by the network operator to generate traffic to monitor the overall network performance, whereas passive probes monitor data flows between different network entities. Because of this, passive probing appears as a possibility to gather information from MNOs' networks.

Capabilities of passive probing are usually achieved by deploying licensed software together with required hardware to the network. There are several possible locations for such probes, most commonly between the BTS/Node-B/eNode-B and BSC/RNC/MME or BSC/RNC/MME and MSC (see figure 2.1). These probes can also query databases, such as the VLR, that store relevant user data. All these methods, however, expect monitoring or data acquisition systems to be "online" in order to constantly gather and

store the acquired data that is temporary in nature. For example, the VLR updates its records every time new information about users' locations becomes available. From the MNOs' point of view, these methods come with a cost of additional network load and might not always be desirable.

Passive probing gives access to data types that usually won't be stored in billing centres and data warehouse systems. This means that besides data that is actively created by the users (e.g., calling), data from network or mobile device-induced processes is also stored (sometimes also known as signalling data). However, it is important to understand that not all information is mandatory for an MNO to store. For example, it is mandatory to store the LAI in the VLR but it is not compulsory to store the CGI together with it.

This kind of data increases the number of events per subscriber considerably, meaning that more resources need to be dedicated for the storage and processing of such data. Pros and cons arise. On the one hand, for instance, increasing the number of records per subscriber helps to minimize diurnal and daily differences in record counts that are the result of subscribers' calling patterns. On the other hand, however, involving system-generated data in statistical analyses might result in high levels of background noise that needs to be addressed in the data processing phase. If it is unknown how the data is generated in the MNOs' networks or what the record types are, it might become increasingly difficult to remove noisy data or to interpret and analyse such data.

The cost of installing the probing systems in the MNOs is usually high and it is not implemented simply for generating statistical indicators for the NSIs. But if the MNOs have implemented some sort of a probing system, it is a good source of data for statistical indicators as it involves much more data compared to "traditionally" collected CDRs from a billing system of data warehouse.

### 2.7.3 General Data Extraction Process

Individual steps of the data extraction process depend heavily on the operational setting agreed between the MNO and the statistical office, mainly on who is responsible for data processing and on where it will be done. If the data will be processed by the MNO or at MNO's premises using a Sandbox-like platform, several extraction steps, like encrypting the data, can be omitted.

By and large, the generic steps are:

1. Data preparation.
2. Data anonymisation.

3. Data encryption.
4. Data transmission.
5. Data archiving.

As stated above, depending on the concrete operational setting, some of these may not be necessary. We give a brief description of each one.

**Data preparation**

The data preparation step typically involves creating a data extraction script to extract necessary data from the storage unit. It should be done by the MNO, most likely with database communication languages such as SQL. The extraction script preferably extracts data automatically at fixed intervals previously agreed upon between the MNO and NSI. This way it is possible to minimize delays that can be caused by the human factor.

The data preparation script should take into account details such as the type of data to be sent (e.g., CDRs, IPDRs or both), time period and attributes that are needed, also the file format and need for anonymization (see next step). The data preparation script also involves some basic data processing steps that the MNO might undertake to provide high-quality data. These steps would involve the removal of non-representative data (e.g. machine-to-machine data) or subscribers that are black-listed for security reasons.

As was stated before, the data preparation step would require the MNO to assign a person to write this script, and, depending on the execution, either to run it at certain times manually or to make sure that automatic processes are working as expected. For example, if there are delays or data are missing, this person should be able to analyse and find why these problems occurred.

**Data anonymisation**

The purpose of the data anonymisation process is customers' privacy protection. During this process, a subscriber's personal identity code can be modified or data can be aggregated to give anonymity to the subjects. Although there are several alternatives to do that, no definitive method currently exists to maintain all the aspects of the data needed for longitudinal analyses (thus longer study time periods) while at the same time making them anonymous to the required extent. Despite the fact that identifying subscribers by the IMSI would be the ideal case, different legislation and MNOs' policies limit this possibility and similar alternatives that can be actually used. In Report 2 of Eurostat et al. (2014), in pages 72-75 the reader will find a detailed analysis of the diverse

alternatives assessing both their degree of protection as well as their suitability and performance for statistical purposes.

In any case, when considering these options during negotiations, the scope of the project where data will be used should be discussed in detail to understand what anonymisation method would be feasible for the NSI as well. Anonymisation (if present) should be part of the data preparation script. If pseudonymous unique codes will be used, the MNO needs to create and maintain hash functions to guarantee the same key and value pairs for the time period agreed upon.

**Data encryption**

Data encryption is a small but important step that is needed when data processing takes place outside MNOs' premises. Its purpose is to ensure that unauthorized parties are not able to read the data in case of a security breach during data transmission (see below).

During data encryption, a key is generated that will define how data will be encrypted into a cipher text. In case of symmetric encryption, the same key will be used to decrypt the data. For asymmetric encryption, a key pair is generated where the public key will be used to encrypt the data and the private key to decrypt it. It is generally advised to use asymmetric encryption if keys will be exchanged over the Internet.

There are several cryptographic libraries available to use for encrypting and decrypting data. Examples include OpenSSL and GNU Privacy Guard, both of which can be used for commercial and non-commercial purposes.

**Data transmission**

During transmission, data will be made available for the NSI either for in-house processing or processing outside MNO's premises (centralised). In the first case, data is generally transported into a database dedicated for this purpose or made available in the form of a data file on a server that the receiving end is able to access. If centralised data processing is used, the easiest ways to transport data are:

- the MNO will dedicate server space to extracted data and make it accessible to the receiving end. The receiving end will connect to the server and transfer the data files to their dedicated server space;
- the receiving end will dedicate server space to extracted data and make it accessible to the MNO. The MNO will connect to the server and transfer the data files.

As can be seen from above, in either case, both parties would need to dedicate some server space to transfer the data.

**Data archiving**

In case of an accidental data loss due to technical problems or any other unforeseeable cause, it is important that historical data could be reacquired and used for recalculation. For the MNOs this simply means the re-extraction of historical (archived) data. However, if data is transmitted to an external processing facility after the extraction, all extracted raw data should be kept in a dedicated storage unit that is not connected to the same unit where the data processing takes place. These archives should follow the legislation concerning data retention and kept only for the period of time allowed by regulations.

### 2.7.4 General Data Extraction Guidelines

As a guiding aid, the following tables will describe the logical order of steps in the process of data extraction. Questions in the table act as guides during a negotiation between an MNO and an NSI with the most common discussion points indicated on two levels. Discussion points present common options that are available.

These guidelines do not provide specific solutions but are a step-by-step technical coverage of all necessary aspects being part of the negotiations with the MNOs.

| Level 1 Discussion Points | Level 2 Discussion Points | Tasks by MNO | Tasks by NSI |
|---|---|---|---|
| **Where can data be processed?** | | | |
| In-house (at MNO's premises) | | 1. Hardware and software setup<br>2. Maintenance<br>3. Hiring data scientist for cleaning and processing the data (if there is no remote connection for the NSI)<br>4. Extraction and processing<br>5. Establishing a data connection to the NSI to deliver indicators | 1. Delivering methodology (and algorithms) to the MNO for producing statistical indicators<br>2. Hiring a data scientist for cleaning and processing the data (if there is a remote connection to the MNO)<br>3. Providing infrastructure to receive indicators |
| Centralized (at NSI's or third trusted party's premises) | | 1. Periodical extraction of data<br>2. Data anonymisation<br>3. Secure data connection to the NSI | 1. Hardware and software setup<br>2. Maintenance<br>3. Hiring a data scientist for cleaning and processing the data<br>4. Establishing a data connection to the MNO to receive data |
| **What is the origin of the data?** | | | |
| Network monitoring (probing) | No existing monitoring system | 1. Hardware and software setup<br>2. Maintenance | Provide the MNO with a list of necessary attributes |
| | Monitoring system exists | 1. Provide the NSI with information about the accessibility of these databases<br>2. Overview of other possible data sources<br>3. Overview of available attributes<br>4. List of restrictions (if needed) | Provide the MNO with a list of necessary attributes |
| Billing centre | | | |
| Customer database | | | |
| Data warehouse | | | |
| OSS | | | |
| Other | | | |

| Level 1 Discussion Points | Level 2 Discussion Points | Tasks by MNO | Tasks by NSI |
|---|---|---|---|
| **What data sources are available?** | | | |
| Domestic | | 1. Provide information about the extent of data available for extraction (time period for each data source)<br>2. Provide information about approximate amount of data (records) | Definition of time period needed for analysis |
| Inbound | | | |
| Outbound | | | |
| Customer database | | | |
| Geographical referencing database | | | |
| **What is the data type coverage?** | | | |
| CDR data | | Provide information about the extent of data available for extraction (time period for each data source) | |
| IPDR data | | | |
| Other | | | |
| **What is the cellular system generation coverage?** | | | |
| 2G | | Provide information about the extent of data available for extraction (time period for each data source) | |
| 3G | | | |
| 4G | | | |

| Level 1 Discussion Points | Level 2 Discussion Points | Tasks by MNO | Tasks by NSI |
|---|---|---|---|
| **What is the record type coverage?** | | | |
| Call | Incoming | Provide list of available record types | Definition of record types needed for analysis |
| | Outgoing | | |
| SMS | Incoming | | |
| | Outgoing | | |
| MMS | Incoming | | |
| | Outgoing | | |
| Internet traffic | | | |
| Location area update (LA change) | | | |
| Location are update (periodic) | Time period | | |
| Other | | | |
| **What is the subscriber coverage?** | | | |
| All subscribers | | | |
| Sample of subscribers | | Sampling | Sample profiling |
| **What is the user coverage by payment type?** | | | |
| Post-paid users | | | |
| Pre-paid users | | | |
| **What is the data coverage by charging system type?** | | | |
| Offline charging | | | |
| Online charging | | | |

| Level 1 Discussion Points | Level 2 Discussion Points | Tasks by MNO | Tasks by NSI |
|---|---|---|---|
| **What attributes are available in the data?** | | | |
| Subscriber identifier | IMSI<br>MSISDN<br>MSRN<br>TMSI/P-TMSI<br>LMSI<br>Other | 1. Provide a list of subscriber identifers<br>2. List of subscriber identifier characteristics | 1. Define expected subscriber identifier characteristics<br>2. Define subscriber identifier characteristics that are not suitable for statistical analysis purposes |
| Location identifier | CGI<br>LAI/TAI/RAI<br>Other | 1. Provide a list of location identifers<br>2. Provide information about spatial accuracy of identifiers | Define spatial accuracy expectations |
| Time attribute | MNO local time zone<br>UTC time zone<br>Destination country local time zone (outbound) | Agree on time zone | Agree on time zone |
| Other attributes | Data type for distinguishing between CDR and IPR data<br>Record types (e.g. outgoing SMS or call,...)<br>Antenna type (2G, 3G, 4G)<br>Cell azimuth<br>Cell sector angle<br>Country code for inbound and outbound data<br>Subscriber type (for M2M exclusion) | | |

| Level 1 Discussion Points | Level 2 Discussion Points | Tasks by MNO | Tasks by NSI |
|---|---|---|---|
| **What are the formats of identifiers?** | | | |
| Subscriber identifier | Standard format (e.g. IMSI) | Agree on format | Agree on format |
| | Integer token | | |
| | Hexadecimal hash | | |
| | Other formats | | |
| Location identifier | Standard CGI | Provide geographical reference file | |
| | CI Integer | | |
| | CI hexadecimal | | |
| | Longitude & Latitude (in decimal) | | |
| | X & Y coordinates (in integer) | Agree on projection | Agree on projection |
| | WKB | | |
| Time attribute | Date-time ISO standard | Agree on format | Agree on format |
| | Other date-time format | | |
| | Unix epoch | | |
| Other attributes | | Agree on format | Agree on format |
| **How will references to countries be handled (inbound and outbound data)?** | | | |
| Extraction MCC from subscriber identifier | | | |
| Separating field with predefined format | ISO 3166 A2 | List possible formats | |
| | MCC | | |
| | PLMN code (MCC+MNC) | | |
| | Country name (not recommended) | | |
| **What is the method for privacy protection?** | | | |
| Data aggregation (in-house only) | | 1. Agree on aggregation level<br>2. Develop aggregation algorithm | Agree on aggregation level |
| Unique pseudonymous code | Permanent | | |
| | Temporary 2 years | 1. Agree on unique pseudonymous code duration<br>2. Develop temporary identifier script<br>3. Provide the same unique pseudonymous code for the duration agreed upon | Agree on unique pseudonymous code duration |
| | Temporary 6 months | | |
| | Temporary 1 month | | |
| | Temporary 1 week | | |
| | Temporary 1 month | | |
| | Temporary 90 minutes | | |

| Level 1 Discussion Points | Level 2 Discussion Points | Tasks by MNO | Tasks by NSI |
|---|---|---|---|
| **How will data be organized in files?** | | | |
| CDR and IPDR | All data sets in separate files | | |
| | Data sets possibly combined into one file | | |
| Domestic, inbound and outbound | All data sets in separate files | | |
| | Data sets possibly combined into one file | | |
| **What are the specifications for data file elements?** | | | |
| Attribute names | | Provide list of attribute names | |
| Attribute types | Decimal number | | |
| | Integer | | |
| | Text | | |
| | Datetime | | |
| File names | | Agree on file names | Agree on file names |
| File type | csv | Agree on delimiters | Agree on delimiters |
| | txt | | |
| | xml | Provide XML schema | |
| | database | Build a database | Build a database |
| **What are the specifications for data extraction script?** | | | |
| Removal of non-representative data | Done by the MNO | Agree on data to be removed | Agree on data to be removed |
| | Done by the NSI during data preprocessing | | |
| Removal of black-listed subscribers | | Remove black-listed subscribers | |

| Level 1 Discussion Points | Level 2 Discussion Points | Tasks by MNO | Tasks by NSI |
|---|---|---|---|
| **How will encryption be handled?** | | | |
| Encryption software | OpenSSL<br>GNU Privacy Guard<br>Other | Agree on encryption method | Agree on encryption method |
| **How often can data be extracted from the storage systems?** | | | |
| Every day<br>Every month<br>Other time interval | | Agree on time interval | Agree on time interval |
| **How should the data be transferred?** | | | |
| File pulled by NSI | | | Develop scheduler |
| File pushed by MNO | | Develop scheduler | |
| Database pulled by NSI | | | Develop scheduler |
| Database pushed by MNO | | Develop scheduler | |
| **Where should raw data be archived?** | | | |
| At MNO's premises | | 1. Hardware and software setup<br>2. Maintenance | |
| At NSI's premises | | | 1. Hardware and software setup<br>2. Maintenance |
| No data archiving | | | |

## 2.8 Costs and skills

Intricately related to the operational setting in which data are generated, it is important for the negotiations regarding access to data to have an idea of the potential costs and necessary staff skills to finally bring access to reality.

Partners of the present work package have almost no experience regarding costs and skills and only in research conditions. Thus we consulted external experts from the Estonian company Positium, where statistics are already being in production in a sustained way (Positium, 2016). Production costs inside MNOs already producing this kind of products are kept under industrial secrecy and cannot be shared with external stakeholders for obvious reasons. In any case, consulted MNOs do agree on stressing the underlying costs associated to any data extraction operation. The issue of data extraction costs, in our view, is central to have access to data for long-term production in a sustained way. It is still an unsolved question.

Exclusively regarding access (processing will be dealt with in the SGA-2 of the present project), there are two major stages where tools and skills are required:

1. development of the system;
2. regular (periodical) data sharing in standard production conditions, including maintenance and data quality assurance.

One-time data extraction and data preparation for transmission should take 1–2 person-months for an MNO. This includes all steps of the general data extraction process.

In order to set up a system that regularly extracts and transmits data, some development is required, which would automate the majority of the processes from the extraction (also possibly some pre-processing) of the mobile phone data. There are no specific additional skills required apart from those presented in table 2.2. However, writing automated scripts for regular extraction, transmission, and checking is required from the MNO and that should take an additional 1–2 person-months. The automated system usually also needs to be bound with the MNO's automated services and the health and quality of the service must be monitored, which might require additional work. For regular data updates, the MNO has to assign specialists who make sure data are automatically transmitted and able to make corrections upon request if there are problems with the data. Depending on the frequency and the service level agreement, this might simply be a specialist who is ready to take a look at the problem and correct it within several days (monthly data updates), or a dedicated team providing 24/7 support who monitor that the real-time data feed is up and running, with very fast reaction and correction periods.

| | Extraction (preparations) | Transmission |
|---|---|---|
| **Hardware** | Data extraction system from MNO's data warehouse. Existing system can be used. | Secure tunnel for transmitting the data from the preparation environment to the processing environment (if this is not in the same location). |
| **Software** | Tools usually included in MNO's data warehouse system. | Data encryption software (available also as open-source) |
| **Skills** | Specialist familiar with the data warehouse system, OSS and any other database or registry of the specific MNO where the data is extracted from. | Knowledge on how to compress, encrypt and transmit large data files. |
| **Person-months** | 0.5 | <1 day |

**Table 2.2** Hardware, software, labour resources and skills to conduct *one-time* data extraction.

| | Extraction (preparations) | Transmission |
|---|---|---|
| **Developing automated system** | 1-4 person-months | 1 week |
| **Maintenance (monthly updates)** | 1 person, part-time, responsible for monitoring the automated process and reacting to errors within several work days. | |
| **Maintenance (daily updates)** | 2–3 trained specialists from the MNO's 24/7 support team who are ready to react and correct errors within hours of reporting the problem. | |

**Table 2.3** Required estimated time for developing the continuous data update system and maintaining it.

Some companies provide platforms for the full chain of data processing, which requires configuration and setup for importing and preparing the data, and also includes the pre-processing and statistical processing of the mobile phone data with ready-made methodology, semi-manual quality assurance tools, as well as integration with existing statistical database systems.

3

# Business guidelines

**Executive summary**

This chapter contains the main findings resulting from the workshop held between the ESS producers and European MNOs. These findings help us understand and disentangle many of the factors behind the difficulties to access mobile phone data by official statistics producers.

Several on-going initiatives of collaboration between NSIs and MNOs were exposed. A round table and a tour-de-table were held with a vivid debate on the different aspects involved in the access to data.

As main relevant issues, these were identified:

- Construct clear use cases to show feasibility and mutual trust.
- Consensus on partnerships outperforming mandatory scenarios.
- Concerns arise when moving from research to production.
- Distributed vs centralised data processing models. Solutions based on development of open algorithms to be applied on secured data kept in data centres (e.g. MNOs premises).
- Regulation on data protection and relationship with regulatory authorities is a big issue at national and European level. More clarity needed.
- Perception by society on the use of these sources. Communication strategy is needed. Transparency for citizens.
- Vicious circle in data access: it is necessary to build detailed case studies and delimit a precise set of data to be requested to MNOs... but some kind of data access is needed in advance for setting up these detailed case studies...
- Relationships with MNOs could be different depending on their current strategies on Big Data. NSIs must take into account these different starting points.
- Quality assurance framework for our users.

A cautious reader, or a statistical officer experienced in negotiating the access to mobile phone data, may rapidly argue that the decision to grant access to data for official statistics purposes is indeed a business decision, not a technical decision.

So, why the long introduction to technicalities about the operational setting of MNOs and the generation of mobile phone data? From our experience, being the mobile telecommunication industry in general and mobile phone (meta)data generation somewhat alien to the official statistics industry, we find it very convenient to know about these technicalities to better understand MNOs' business position when negotiating with them. This can help official statisticians find their way to reach an agreement.

As a milestone in the SGA-1 of the present ESSnet, the work package on mobile phone data organised an international workshop at Luxembourg gathering representatives of both industries, in particular statistical offices such as ESS NSIs and Eurostat and European MNOs. International organizations were also invited. During the meeting, there was an excellent participative atmosphere in which attendants expressed their view and position about potential public-private partnerships for mobile phone data for use in official statistics production.

A detailed account of the celebration of this meeting will provide an ample insight into the diverse aspects related to the access to mobile phone data for official statistics production from different angles.

## 3.1 Context and actors

With the goal of covering as many positions as possible the workshop was divided into four sessions. In session I, the presentation of both context and actors regarding the use of mobile phone data in official statistics production was conducted.

Mobile phone data appears as one of the most promising Big Data sources. However, the official statistics industry has to face multiple challenges in the road to investigate the potentialities of these data for the production of European and national statistics. Among them, we can mention the construction of a collaborative and strategic partnership with the private sector, and to find a space of understanding in which the interests, concerns and needs of official statistics and private companies can coexist.

The mission and values of Eurostat and the European Statistical System (ESS) gives a natural supporting framework for the potential use of mobile phone data in the production of official statistics. The key element of European official statistics production is the Vision 2020, a strategy for the modernization of the ESS based on five key areas in which

common action is needed in order for European statistics to be "fit for the future". They can be summarized as (i) meeting users' requirements, (ii) facing methodological and technological challenges and (iii) innovating its offer in terms of products and services. Data revolution brings opportunities and challenges, and the ESS has to respond to this change of paradigm using all the relevant sources to produce knowledge preserving always our high commitment to quality. To be successful facing all these challenges, the statistical community cannot work alone, we need to build partnerships and to seek strong political support and investment. Thus, aiming at the use of mobile phone data under a collaborative partnership with private sector stands up as a key ingredient in the immediate future in the strategy of the ESS.

The origin of the focus on these new Big Data sources, including mobile phone data, is rooted in the increasing digitalisation of human activity and the great opportunities this brings along for the production of official statistics.

Indeed, the digital footprint that humans left behind can be used to improve statistical production. The Scheveningen Memorandum on Big Data, agreed by the ESS in 2013, called for the submission of an action plan and roadmap on Big Data, adopted a year after, in September 2014. The main aspects of this action plan are the integration of a Big Data strategy for statistics into an overall government strategy, the need for skills, the collaboration at European and global level, the need for methodological developments, quality assessment and IT, the search for partnerships between different stakeholders (government, academics, private sector), communication and the protection of privacy and personal data (legal and ethical issues).

As stated in the preceding chapter, in 2012-2014 Eurostat carried out a feasibility study on mobile positioning data for tourism statistics. From this study an analysis of the strengths and obstacles in the use of Big Data sources was derived. Among the opportunities a better density in terms of space and time, as well as higher coverage and observation of behavior can be mentioned. On the other hand, problems are encountered in issues as selectivity and bias, lack of additional information to study socio-economic characteristics, integration of different sources and their sustainability.

A first set of challenges for the future refers to a common and scientific approach of the ESS to the use of Big Data sources based on cooperation and exchange of best practices. To address this issue, Eurostat is launching a series of pilot projects under the framework of the present ESSnet on Big Data. These projects are an important pillar of the Big Data activities in the ESS in the coming years and should pave the way towards the integration of Big Data sources into official statistics production.

Secondly, new skills and IT infrastructure are also key to successfully move towards

the new sources. To this aim concrete actions are ongoing, for instance the "sandbox" environment for Big Data experiments hosted by the Irish Central Statistics Office in a cooperation between Eurostat and UNECE, among others. On the other hand, the training strategy is under the umbrella of the European Statistics Training Programme. The courses are focused on sources and tools to assess, use and explore Big Data. Political and regulatory framework are also important areas to work on, especially related to the access to data. The discussion is not limited to the entry but should include a long term vision to guarantee a continuity of access.

Other important challenges are data security and privacy concerns or how to find a sustainable business model for Big Data in official statistics.

In summary, the European official statistics community must go on working on Big Data projects, trying to establish partnerships, exploring possibilities of combining sources, and finding new uses and applications.

## 3.2 On-going initiatives

During negotiations, more often than not it appears the issue of on-going projects of similar characteristics. Within business strategies, MNOs are usually more comfortable when similar projects are already going on. During the workshop, we presented three of such projects. Full details are not needed for the negotiations, but it is fairly positive to have an idea of the overall description of these projects to be aware of some of its possible implications.

### 3.2.1 Initiative 1: Mining mobile phone data to recognise urban areas

INSEE gave a presentation on an on-going joint project with Orange Labs Sense and Eurostat . They are currently working on CDR (call detail record) data sets from 2007 (May-October), approved by national authorities to be used for statistical purposes. The project includes two case studies: (i) urban areas detection and (ii) home detection. These lead to two use cases, at individual level (done in situ at Orange Labs) to use individual trajectories to identify residence, and at aggregate level (done by INSEE) to classify areas. The objective is to infer area type from mobile phone patterns (residential vs. working) and also to compare a new source with a reliable reference (census data from 2008).

The processing for urban area prediction was made in three steps:

1. To split data in training sample and testing sample. As information is very spatial correlated it has to be first decorrelated.

2. To estimate a prediction of antenna's classification on the training sample and calibrate the algorithm (using several methods and different metrics). Benchmark mobile phone data using sociodemographic variables coming from official sources (census 2008).

3. To analyse the results. They show that accuracy is reasonably good, although the prediction can vary among classes.

The conclusions of this pilot study are the following:

1. Joint work between NSI and MNO is crucial to develop expertise in the NSI and to evaluate the potential use of this source in official statistics. This partnership is beneficial to the MNO as a way of validating their data for other business purposes by having the quality stamp of official statistics.

2. Aggregated mobile phone data are compliant with sociodemographic information used to classify area, and enough to satisfy information needs.

3. Further work is needed to replicate the study on more recent data, in order to control what is going on over time and to explore the stability of the results.

4. Further analysis is also necessary at a more detailed level.

### 3.2.2 Initiative 2: In-house (distributed) and trusted third party (centralised) approach for processing mobile phone data: pros, cons, and market share conflict points

Positium's contribution showed the options for processing mobile phone data based in their collaboration with Statistics Estonia and other countries.

Generally, when talking about mobile data processing three phases can be identified: (i) extraction, (ii) processing, and (iii) aggregating results, and they can be tackled from three different options:

1. Distributed (in-house) processing. Data are handled by MNOs from the extraction to aggregation and calculation of indicators, while NSIs can combine the results from several data providers. Under this model privacy protection aspects are solved and less skills are needed for NSIs. Moreover, MNOs can use the data also for their commercial aims. In general terms, this option is less expensive for NSIs in terms of direct costs. On the other side, NSIs do not have access on raw data and no control over the methodology nor the processing that might be different in every case. Also the entailed burden on MNOs requiring compensation must be mentioned.

2. Centralized processing. Extraction of the data is done by MNOs and raw data (anonymised or semi-anonymised) are transmitted to NSIs who will be responsible for the full processing chain. This guarantees the use of one single methodology, as well as trusted and controlled results. NSIs have a complete control over methodology and algorithms and there are no burden on MNOs further than the costs of extraction. The main disadvantage lays on the privacy protection that legally prevents this option in many countries. Additionally, technical resources and skilled employees (data scientists) are required. Overall, this option is directly more expensive for NSIs.

3. Centralized processing trusted in a third party. MNOs extract the data and are transmitted to a trusted third party (government organisation, university, company,...). In principle this option would guarantee one methodology and almost controlled results provided that all MNOs transmit their data to the same third party (otherwise the complexity of the business model would be evident, even higher than in any other options). There would be no burden on MNOs and less need for new skills for NSIs. The highest risks are the unknown methodology, the privacy protection and the higher degree of complexity of the business model as more stakeholders are involved in it.

Technically all these options are possible, thus the best depends on the business model and national legal conditions.

Another element to take into account is the market conflict that could exist between public interests and the private sector. To illustrate this point, Positium showed some examples where the market for data could overlap and mentioned the need to agree on the division between markets.

According to Positium (WP5, 2016) the top barriers to Big Data adoption in Official Statistics (as of September 2016) are the lack of funding, not enough policy-makers engagement, difficulties on identifying key partners and get access to data, as well as to assure new skills, capacity building and technology.

However, in the European Union the biggest obstacle is regulation and privacy protection that could require changes in European and national laws. Still, from Positium's experience, this in-house processing would be acceptable in several countries that if done by MNOs and identifiable data were deleted after the results.

In any case, to overcome all these obstacles, it is important to make the actual return of investment (ROI) for the society in the long term more visible.

During the discussion many considerations and concerns about the business model arose. NSIs are concerned about issues such as solving the representativity of data, especially in a decentralized model in which there is no control on methodology, and considering that many MNOs do not see the importance of representativity because their customers do not ask about that.

Moreover, MNOs stressed the need to resolve issues such as security, protection of privacy and reputational risk of customers perceiving negatively the transfer of their personal data. They also highlighted the fact that MNOs can be in different situations: some of them can have a business strategy on Big Data, some can be developing a commercial line and others have not explored yet their use. For MNOs that have carried out a large investment in technology, innovation and human capital to learn how to extract value from Big Data is very hard to give away or jeopardize the monetization of that strategic investment. Not to mention the strong competition inside the sector in which sharing or generalizing this knowledge, when the level of development between the companies is unequal, can seriously harm the interests of those companies that already have a Big Data product on the market.

All these questions have no clear answer at this time, and they can only be resolved by trying to move forward.

### 3.2.3 Initiative 3: Value from mobile phone data: a mutually beneficial partnership between a network operator and a statistical office

Proximus presented their collaboration experience with Statistics Belgium as a concrete example of a jointly positive partnership.

First, the barriers and problems identified by Proximus were fairly much the same discussed in the preceding section, mainly privacy and confidentiality concerns. Big Data are still in an innovation stage that requires a great effort of investment for the MNOs to be able to develop a commercial product or service based on mobile phone data. This makes them more sensitive to the exposure of knowledge competition. From NSIs side, the constraints focus on the data access, the lack of Big Data experience and IT infrastructure and the absence of a specific legal framework.

In this context, joining forces can bring important contributions. MNOs can provide data, metadata, infrastructure and technical expertise that could make NSIs progress in the production of faster statistics, with more detail and better validations, coverage and concepts, as well as to reduce response burden and costs. On the other hand, NSIs can offer to MNOs geocoded statistical datasets, statistical and methodological learning to apply in commercial products, domain and subject expertise, official quality stamp

and an opportunity to improve their credibility and corporate social responsibility (CSR).

Proximus, Statistics Belgium and Eurostat tested this approach with a concrete cooperation case aiming to assess the quality of mobile phone data as a source of statistics. From this experience they had learnt the following lessons:

- Mobile phone data is relevant for official statistical production.
- There is a need for close cooperation between public and private sectors to reach a common understanding of what is (and is not) in the data.
- The need for complementary skills and resources.
- Combination of datasets is a necessity, protecting data privacy.
- This kind of partnership gives MNOs external and internal visibility.
- It is positive to start with small and pragmatic projects.
- To get a trusted and solid cooperation is important to focus on mutual benefit, there are many spaces for a win-win.
- The value of this pilot study was widely recognised, although some MNOs expressed that this kind experiments are acceptable in a pilot environment but many concerns appear when we think on moving to production.

### 3.2.4 Round table: lessons from these experiences

During the workshop we organized a round table to collect as many viewpoints as possible. Eurostat acted as moderator and Statistics Belgium, Proximus, INSEE, Orange Labs, Statistics Finland and Telenor Norway as panellists.

Eurostat opened the round table by bringing up two questions that acted as a common thread for the discussion:

1. What are the good lessons (opportunities) and lessons to avoid (risks) in order to build a more solid partnership?
2. Concerning data access level, what are the general lessons learned and guidelines to follow?

The main relevant comments shared by the panellists and other participants are listed below:

1. An outstanding priority placed on the implementation and development of good and clear use cases to show real usefulness, feasibility and mutual trust was agreed by everybody. We must learn by doing, and a good starting point is to begin from something small that allows us to take pragmatic and concrete decisions. To this aim INSEE-Orange Labs, for example, have worked with old and aggregated data (2007) to avoid confidentiality problems and to help to create that trust.

   Related to this, some MNOs called for greater clarity and precision over NSIs' needs and the data that may be required on a continuous base for production. However, NSIs find difficult to pinpoint without having accessed the data, hence the importance of developing practical case studies.

2. There was a consensus on partnerships outperforming mandatory scenarios. One of the most important aspects is to build trust between the public sector and private companies and the most effective way is to do it within a collaborative environment. Empathy has to lead the manner to approach MNOs. Public interest cannot disregard the investment made by MNOs nor their fair right to take commercial advantage from Big Data. Efforts have to be made in order to assure a common interest, so that NSIs could also offer to MNOs added value (statistical methodological knowledge, corporate image...).

3. A need to find and to agree on the boundaries of markets to avoid a business conflict between public and commercial interests was recognised. Maybe the key is not to define specific domains of customers for NSIs and MNOs but to come to an understanding on the granularity of the information disseminated freely by official statistics. That is the level where the line has to be drawn to find the win-win situation.

4. Concerns arise when moving from research to production. This aspect is closely related with the two last points. Cooperation with Orange Lab, for instance, was possible because from the beginning the project was shown as a small and limited research study not production oriented. Moreover, the use of Big Data was not at that moment at the business core of the company and therefore there was space for a project of that characteristics. The question of moving to industrialised process is totally different, not yet solved and requires a very careful analysis. There are different levels of maturity and the stage of production is seen by the time being in the long term.

5. Distributed vs. centralized models. To process mobile phone data for the production of official statistics two extreme situations can be possibly considered. On the one hand, data can be kept in their original information systems at MNOs' data centres and processed there only to bring out compiled partial aggregates to NSIs.

On the other hand, data can be securely transmitted from MNOs to NSIs (or to a trusted third-party) where they all are combined and jointly processed. In the former case, data privacy is prioritised whereas in the latter case traceability and auditability of the whole process is preferred. Solutions based on the development of open algorithms (OPAL) that could also be used for commercial purposes appear as an interesting candidate. This could be a good approach satisfying NSIs needs for auditable methodology. The bottom line is not to throw data onto algorithms but open algorithms onto undisclosed data.

6. Regulation on data protection and relationships with regulatory authorities is a big issue at national and European level. More clarity is needed to define the (new) complex legal framework where a convergence of business sector regulations (that rules the use of information produced by MNOs economic activity), statistics regulations (that gives public statistics producers the right to access data) and data protection regulations (that generally protects citizens' sensitive data) is already evident.

7. Perception of the society on the use of these sources. A common communication strategy is needed for avoiding public opinion to reject the statistical use of mobile phone data.

## 3.3 Core issues regarding access

In the third session we present both an NSI's and an MNO's perspective on the core issues regarding the access to mobile phone data for official statistics purposes.

### 3.3.1 Official statistics perspective

Statistics Spain (INE) presented this topic. Firstly, the global framework was addressed, mentioning the UN Fundamental Principles of Official Statistics, UN Recommendations for Access to Data from Private Organizations and the European Statistical Code of Practice. The issue of confidentiality and privacy is on the base of these principles, and they recognise that private companies have to provide data, although a proportionality and fair balance is demanded between information needs and data required. The cost and effort of providing data access must be reasonable compared to the expected public benefits. For negotiations it is an important point to clearly show that privacy and confidentiality is in the DNA of official statistics producers.

Next, Statistics Spain (INE) pointed out some hints about different characteristics of the MNOs and aspects related to legal requirements, access conditions and data characteristics (following indeed the contents of deliverable 1.1).

Concerning official statistics position towards the use and integration of Big Data in statistical production, Statistics Spain (INE) expressed the following:

- Mobile phone data have a clear public interest for official statistics purposes.
- Official statistics producers are aware of and sensible to the complexity of factors involved in the access to data (with the need to figure out how they interrelate among them): specific characteristics of each MNO, unclear legal regulations, different access conditions (in-situ or transmitted, research vs production,...), confidentiality and privacy, public perception, collision of interests,...(see deliverable 1.1).
- To tackle these issues and develop case studies, NSIs need preliminary access to data.
- Partnerships/joint ventures are the optimal approach to meet public and private interests.
- Common solutions will help to guarantee the privacy and confidentiality of data, as well as a sustained access to data source to assure statistical production.

Once again, MNOs noted that it was necessary to define what kind of data sets and variables are expected to be collected by the NSIs because on that depends the effort that has to be done to extract and store the data. NSIs replied that fixing a priori the amount, characteristics and frequency of data is not possible, research and case studies are needed in advance. In any case, all participants agree on the relevance of developing a model based on a fair balance of the benefits, needs and costs from both sides.

NSIs also mentioned that data integration was an area where Official Statistics could bring very much knowledge and added value on commercial production, both methodologically and providing linked aggregated data from official statistics and other administrative sources. The role of NSIs as a potential trusted party where to aggregate and integrate data from diverse MNOs to control market share bias was mentioned as an interesting option.

### 3.3.2 MNOs' perspective

Telefonica, as a renowned company in European telecoms market and user and developer of Big Data solutions, presented this issue and gave their MNO's perspective.

Telefonica is a data driven company, and Big Data are one of the strategic lines fixed for the future and to which all the company is committed with. The structure of the company itself has been changed for this goal (bringing with it also a cultural change)

and important acquisitions and investments have been made (staff profile, IT,...) to reinforce their capabilities and offer.

At present, Telefonica is using Big Data technologies both for improving internal processes and increasing offer to customers.

Next, the company briefly explained their market product, SmartSteps. SmartSteps uses active and passive events and defines two types of elements, namely, a "stay" (when the mobile user is at the same location at least 30 minutes) and a "journey" (in case any change of location in less than 30 minutes happens).

The starting and ending location destination are identified and stored. Also, each event is related to a user ID, a time stamp and a geographical location.

First, data are anonymized and put into the system to be treated (data storage and formatted), then data are aggregated per cell (no individual is identifiable) and finally extrapolated by applying algorithms to represent the entire population. A set of socio-demographics variables to evaluate the data and the accuracy for the different studies is also used. To guarantee privacy and confidentiality, a minimum level of anonymization and aggregation is fixed in 15 users.

In the framework of the ESSnet on Big Data project Telefonica and Statistics Spain (INE) have been in contact to explore the possibilities of accessing mobile phone data. The idea was to develop a case study to investigate, in a specific statistical domain, the feasibility of using this information source in the official statistics production. The research would be to concentrate on the study of human mobility using the events generated in the mobile network.

Telefonica's proposal to access data is based on SmartSteps product and would comprise one natural year of data for the whole country. The company would be the owner of the database and would provide supporting infrastructure allowing INE to access the data, build the analytics and export the results.

The main obstacle to reach a final agreement was the potential far-reaching consequences for the whole ESS and the Spanish National Statistical System of assuming the noticeable costs associated with such a platform, especially when promoting these analyses from research to production.

After debate, the general feeling of participants was that we have a complicated way to go before making a public-private partnership for accessing mobile phone come true. There are many issues to be solved, especially if want to move beyond the field

of research and analysis. The question is to start working together in a concrete and small pilot study trying to obtain results that may shed light on a future model for public-private collaboration.

### 3.3.3 Exchange of views

To have a further exchange of views, the workshop also organised a tour-de-table, now moderated by the European Commission DG Connect. The main topics discussed by the attendants were the following:

1. How to break the vicious circle in data access: it is necessary to build detailed case studies and delimit a precise set of data to ask to MNOs,...but some kind of data access is needed for setting up these detailed case studies...Although the participants had different opinions, there was a general agreement on giving priority to case/pilot studies.

2. Relationships with MNOs could be different depending on their concrete situation and current strategies on Big Data. NSIs must take into account these different starting points, particularly when trying to solve the question research vs production. An MNO having invested to develop a business line around the statistical exploitation of their data does not face the same situation as an MNO not having this strategy on the core of their business.

3. In this sense, it is capital to show that there is no collision of interests between the dissemination of official statistics based on mobile phone data and commercial products and services based on the same data. Not only should there be ample room for both businesses but also they could benefit from synergies since official data are of great complementary value for these commercial solutions. The limits of the market could be associated to frequency and geography granularity, meaning that at certain high level official statistics would provide free data but at further detailed level MNOs would have assured their market. This approach could even promote commerciality. Participants agreed that this issue would have to be worked out in detail.

4. Many different aspects and connected among them. We have to identify and understand all these interrelationships. For example, legal environment is still a grey area and MNOs ask for more legal clarity, producing guidelines on data protection.

5. A public communication strategy is needed to explain and convince the society of the benefits and guarantees of the use of Big Data in official statistics. Perceptions will need to be analyzed. Citizens are highly sensitive concerning their personal and private data and their potential use by public administration, especially when

these data may come out of the MNOs' information database. A general negative public opinion about MNOs transferring data to NSIs can affect their decision to collaborate.

6. It is important to find ways to motivate MNOs cooperation, to offer them incentives for collaboration: improve and increase public image and visibility for their corporate social responsibility, define markets (granularity), return of information, gain new knowledge from official statistics that can be used for commercial products, validation or quality assessment using data from official statistics (e.g. geocoded population registers), overcome market share bias by letting NSIs integrate and aggregate data from diverse MNOs... We also have to be aware that finding motivation is easier for MNOs that have not already invested great resources in a Big Data strategy and have not developed this competence yet.

7. A quality assurance framework for our users is also relevant, and both, NSIs and MNOs should have interest on it. Quality indicators and framework (quality stamp) can be more easily reached by collaboration.

## 3.4 Conclusions for the future

It is to be highlighted the impact of events like the referred workshop, which, from a constructive approach, help the dialogue and understanding of the needs and interests of the public and private sectors in the field of Big Data.

We all agreed on the relevance of the many issues discussed during these two days and the importance of maintaining the bridges open and active for communication. Options will be explored to keep the dialogue and communication between NSIs and MNOs in order to progress, particularly, in one of the priorities identified repeatedly in different interventions: building new case studies.

4

# Experiences in the ESSnet on Big Data

**Executive summary**

This chapter collects guidelines resulting from WP5 partners' own experiences in negotiating access to mobile phone data for the present ESSnet.

As main guidelines, we have identified the following:

- See/show the window of opportunity of building up a partnership between an NSI and an MNO.
- Get the right people committing their organizations with technical skills and competence to build up the partnership.
- Show empathy and value arising from the NSI's contribution to this partnership.
- Show absolute guarantees of confidentiality and privacy protection.
- Show the limits of producing statistical outputs with no collaboration in contrast to the combination of data sources and methodologies.
- Be aware of the complexity of data extraction and the implications in cost extraction and professional skills.
- What data? Mobile phone data for statistical exploitation do not exist in a mobile telecommunication network and a concrete specification must be formulated.
- Define a concrete small research project.
- Be attentive to legal issues.
- Analyse costs.
- Be transparent.

The top priority of work package on mobile phone data of the present ESSnet during the SGA-1, as stated in deliverable 1.1, has been the individual negotiations of each country member of the work package with corresponding national MNOs to get access to data to conduct the following phases of the research plan during the SGA-2.

In this chapter, we summarised the main points arisen during these negotiations which may serve as a guidance for other ESS partners. Needless to say, these are **from** the official statistics perspective and **for** official statistics producers.

## 4.1 A quick overview

As a first important conclusion from our experiences, it must be stated that no golden rule for succeeding in getting access to data can be given. The situation (thus the points arising during negotiations) are not only different from country to country but also different from company to company within the same country. The situation is thus clearly MNO-specific although conditioned by national factors as the legal framework or the current national mobile telecommunication market.

Secondly, the many diverse factors included in deliverable 1.1 (characteristics of the MNO, legal requirements, access conditions, data characteristics, and other aspects) are intricately entangled in the actual situation. For example, legal requirements may settle little space for discussing several access conditions or cost compensation may put a limit on the coverage of data.

We have detected three extreme poles regarding the characteristics of MNOs. On the one hand, you may find companies clearly interested in Big Data (their mobile phone metadata) as a commercial product. Here two extremal points can be recognised, namely (i) companies having already invested in this business line and (ii) companies beginning their actions and potentially searching for partners or searching for an optimal similar course of action to statistically exploit their data. On the other hand, there still exist companies whose strategic business plan does not include this statistical exploitation. Depending on the actual situation, the rest of aspects will be more or less important and entangled.

In our negotiations, we have encountered the three situations driving us to different aspects of the access. When talking to companies with a Big Data business line division, costs, privacy of their customers, protection of their investment, and legal support, to mention the most relevant, immediately arise as issues at stake. When talking to companies initiating their steps in the exploitation of their data, privacy of their customers and legal support are still issues at stake but the search for profitable business cases

also adds to the discussion. When no interest in Big Data is shown by the company, it is practically impossible to initiate negotiations.

When having initiated contacts with MNOs, we, members of the WP5, have also detected differences within the particular representative or departments with which the talks are carried out. As in the official statistics industry (still having different approaches and sensibilities towards the use of Big Data), in MNOs as well we have detected different approaches and sensibilities. MNOs are large companies where diverse professional profiles (researchers, legal specialists, IT specialists, marketing and sales experts, telecommunication engineers, ...) show different degrees of risk aversion thus either clearly recognising in this public-private partnership an excellent business opportunity to increase the value of their Big Data products or on the contrary demanding a stronger legal support and investment protection which will apparently avoid potential conflicts. In all cases we have approached the question of granting access to data in a collaborative partnership to produce high-quality statistical products. All negotiations having granted us access to data have been conducted in these lines.

From the legal point of view, there exists in all cases an interplay between the European and national legal frameworks. Guidelines in this respect are clearly country specific. However, there is always a common factor: the corresponding national Data Protection Authority (DPA). In all cases, a statement by the DPA has been reached to show the legal support to the access by NSIs according to the corresponding National Statistical Act and related legislation and guaranteeing the protection of privacy of all citizens. In this statement the interplay between the European and national laws plays a relevant role solved by the DPA itself.

By and large, in Big Data sources four main legal issues arise to be tackled, namely (i) access to personal data, (ii) copyright, (iii) public databases protection, (iv) confidentiality. Firstly, as of this writing, the definition of personal data and the legal European framework for their protection is settled in the Data Protection Directive (Directive 95/46/EC), complemented by the corresponding national legislation in each Member Country. In the European realm a priori very little legal obstacles are found to the collection of these data in general, although specific telecom legislation mandating a limited retention and subsequent deletion may stand as a limiting factor for NSIs to collect these data. This conflict between the statistical and telecom legislation must be analysed in each national case. Secondly, mobile phone data are not protected by copyright for their processing by NSIs, however methodologies, protocols, procedures, algorithms or certain operations within MNOs to store, extract or process their data may want to be kept under copyright, intellectual propriety rights or industrial secrecy. This is especially critical in the mobile telecommunication industry, where competitiveness is voracious. This could cause a conflict with the auditability of the production process of

official statistics. This must be solved on a case by case setting reaching a contractual agreement. Thirdly, mobile phone data sets do not constitute a public database and the access thus must be reached under private agreement. Finally, confidentiality of mobile phone data must be warranted in the same terms as any other data possibly identifying individual people. It is remarkable that MNOs are not usually aware of the commitment of NSIs to protect the identity of official data providers (respondents). It is recommended to make it explicit by noting that not only does a strong legally binding framework exist but also that there is statistical methodology explicitly developed and implemented to this purpose in the official statistics production.

As of this writing, we must mention that a European initiative within the ESS has been clearly recognised by the European Statistical System Committee as a need to pave the way for having access to data in private hands. Some works have very recently begun to analyse in close detail the potential actions to arrive at better conditions to access these new data sources.

A specific mention must be made to the equal treatment that must be given to all MNOs from an NSI as a public institution. Apart from the difficulties arising in reaching bilateral agreements, it must be taken into account that all MNOs must receive a similar treatment in our request for data. Public tenders appear as a clear option but given the diverse technical and analytical capabilities of MNOs it may be difficult to find an equal-chance balance in the terms of the tenders. In some NSIs, approval of public tenders must be previously deliberately by an internal ethics committee. Notice that a new potential vicious circle may also appear here: there may arise the argument that it is not ethical to demand such an amount of private and sensitive personal data to produce official statistics (some other data sources may be enough), but to know whether there is a real public benefit at least some proof of concept or empirical assessment (thus accessing the data) must be undertaken to take an informed decision.

The actual operational access conditions are strongly conditioned by the former legal framework and the statement by the DPA. The level of intricacy here is clear. If either raw or preprocessed microdata are to be accessed, then evident measures to protect privacy are considered. If an in-house model (within MNO's premises) is considered, then time and geospatial attributes do not need to be coarse-grained or aggregated. If a transmission model (to NSI's information systems), then either time or geospatial attributes (or both) are coarse-grained to reduce the level of granularity thus impeding personal identifications. In all cases, anonymised IDs are used. Complementarily, old data (say, from a decade ago) are used also to protect privacy.

In all cases work package members are working under research conditions, so considerations for long-term standard production conditions have only slightly appeared in the

negotiations, especially regarding the access to data from all MNOs. For the long-term access more work is needed since details have still to be discussed (e.g. extraction costs).

It is important to recognise from the very beginning of any negotiation a common feature bringing up a crucial source of tension with the traditional data provision principles for official statistics production. MNOs, as companies in a fiercely competitive market, are strongly profit-oriented. This is in apparent conflict with the golden principle of data provision by society for official statistics production at no cost (on Big Data, 2016). The costs associated to data extraction from the complex cellular networks of mobile telecommunication (see section 2.8) will certainly arise in the negotiations. In this sense NSIs should make very clear their positive contribution to this partnership coming from their statistical expertise and their knowledge and use of official data (e.g. of geocoded data sets with rich sociodemographic information). An elementary remark in this respect is the question of inference from the mobile phone data set in relation to the entire population of analysis. Techniques in register-based statistics already known to official statisticians may be of great help here.

Furthermore, in those cases where MNOs have already invested in Big Data commercial products, the protection of their intellectual property rights (not of data but of their tools) will also arise as a concern, especially given the competitiveness of this market. NSIs cannot appear as a vector transmitting this kind of information among different competitors. NSIs must show a clear commitment in this respect.

The combination of mobile phone data with official data, either in microdata or aggregated form, should not only be considered from an operational point of view but also and even more importantly as a relevant input from NSIs in this partnership. Geocoded data sets enriched with sociodemographic variables are commonly composed and used in NSIs. These can be used to enrich mobile data set (not meaning disclosing individual official microdata) and to make it possible to potentially use richer statistical methods to arrive at high-quality outputs (to be undertaken in SGA-2).

The discrimination between accessing data for research, and especially with a concrete research output (not for research in generic conditions *sine die*) or accessing data in a long-term sustained way for standard production is clearly important. In all cases of the work package members having succeeded to reach an agreement in the SGA-1, data extraction has been provided free of charge at no cost for this research project. Details for long-term production have still to be worked out. Due to the absence of precedents, NSIs currently lack expertise to assess data extraction cost amounts within the usual official statistics production budgets, at least at the same level as which NSIs can actually value data collection operations for their subcontracting with specialised companies. As a generic guideline, this must be indeed one of the outputs of the partnership at the

research stage.

As already mentioned, extraction costs will certainly appear in any negotiation. This is an extremely controversial issue, especially regarding the traditional principle of data provision at no cost (on Big Data, 2016). Furthermore, in practice it is highly entangled with some other aspects of data access. It must be clearly stated that costs are considered regarding extraction operations and never for the data themselves. Indeed this distinction is clearly made in on Big Data (2016).

Costs are intricately entangled with the nested cellular structure of the network and the events originating the data. It is not the same feeding your data sets with events at the BSS level which are not permanently stored in the system (see section 2.4.1) as feeding it with data generated exclusively in the billing system (see section 2.5). Furthermore, as deduced from the description of the complex cellular structure of mobile networks, part of the data extraction process must be carried out by the MNOs themselves, which must thus incorporate this process into their production (for long-term conditions) with minimal disruption of their usual processes (accessing and updating internal databases and transmission channels). Being clearly profit-oriented, the data extraction process must be carefully designed as a cost-effective process, much in the same way as the goals posed by the European Statistics Code of Practice for the statistical production process (ESS, 2011). For this reason, as a generic guideline, we find it more convenient to start with a small project with realistic objectives allowing us, among other things, to fine-tune the execution of data extraction and other related tasks.

Characteristics of requested data are another important piece of negotiations, usually arising in the form of a simple question by MNOs: "what data do you mean by mobile phone data?" We have seen the amount of databases included in a mobile network and diverse type of data possibly captured from diverse events at different levels (BSS, NSS, OSS, billing system, ...). According to deliverable 1.1, data can be raw or pre-processed microdata (at subscribers' individual level) or aggregated data. The form of data requested and possibly accessed will depend on various factors. The access to microdata is in general preferred only when processing and aggregation is conducted in MNOs' premises, since it solves issues related to privacy, confidentiality, and legal concerns. When microdata are transmitted outside the MNOs' premises, they are strongly coarse-grained and anonymised preferably according to the DPA's statement (see above). When only aggregated data are agreed to be transmitted out of MNOs' information systems, then not only extraction but aggregation costs may appear in the discussion. As a generic guideline, beginning with a concrete small business case will help both parties (MNOs and NSIs) assess all these factors thus possibly calibrating actual risks and costs to progressively scale up to bigger projects.

In this sense, it may appear wise to begin with a partial coverage of the national territory (also depending on the size of the country) and restricted to a manageable time frame (say, a few weeks). The complementation with additional data attributes like sociodemographic variables from the subscription contracts and technical data about events generating the data can be debated according to legal and privacy concerns. In the specific case of (foreign) roamers, depending on the concrete output for the project (e.g. for tourism statistics), this information must be clearly included in the negotiation.

In the case of MNOs interested in commercialising statistical outputs, more often than not the apparent conflict of interests will arise under the following argument. If MNOs grant access to their data for official statistics purposes, their potential market will disappear since clients will use mobile phone data-based official statistics for free. Two counterarguments can be given. On the one hand, the situation is similar to that of those consultancy firms conducting and selling market analysis studies and reports in the private sector. There exists a consolidated market around this activity even in many cases positively using current official statistics for their business. This is not a clash but a symbiotic relation between the public and private sectors. On the other hand, complementarily the outputs of public statistical operations are fixed by National Statistical Plans and/or the European Statistical Plan, including their scope (level of time and spatial granularity, among others). Concrete interests of concrete private firms are not the goal of these Plans. For example, a sociodemographic or mobility study around selling premises of a supermarket retailer chain are not the goal of these Plans, and rather on the contrary it may appear as a clear statistical output for the private market. The benefit from using data enriched with official data and (partially) using official statistics methodology is obvious in this case. The public sector is focused on providing statistics for policy-making. As a guideline, again the consideration of a concrete business case within the collaboration will help both MNOs and NSIs assess where the frontier between private and public interests lies.

To end up this section, the question of public opinion and customers' concerns about the use of their data must be jointly considered. As debated during the workshop at Luxembourg (see chapter 3), a common communication strategy should be elaborated in which with full transparency the scope of the use of customers'/citizens' data are clearly informed clearly providing the arguments for the legal support (especially from the DPA) and the technical solutions giving to any privacy concern thus avoiding identification risks.

## 4.2 Guidelines: a bullet list

We collect our proposed guidelines as a quick bullet list:

1. See/show the window of opportunity
   NSIs willing to explore mobile phone data as a new source and MNOs having invested in data and looking for their statistical exploitation should find in a partnership an optimal strategy.

2. Get the right people
   Both from NSIs and MNOs it is important to find and assemble those people possibly committing their organizations and those with technical skills and competence to make the partnership a reality.

3. Show empathy/value
   As a *joint* venture statistical officers should understand the competitive market environment in which MNOs operate. They are clearly profit-oriented companies. Make explicit the value of NSIs' contribution to this partnership.

4. Show absolute guarantees of confidentiality and privacy protection
   As statistical officers, clearly show the contrasted firm commitment of the whole official statistics industry to protect privacy and ensure confidentiality of any kind of individual data. This must also include the protection of the know-hows of each company.

5. Show limits
   Show the limits of producing statistical outputs in their own. NSIs need MNOs to access mobile phone data and MNOs need NSIs to produce high-quality statistical outputs by using their statistical expertise and by enriching mobile phone data sets.

6. Be aware of the complexity of data extraction
   Mobile phone data are not a closed entity in the companies. They can be generated in many forms from the *events* between mobile devices and antennae. It is not a matter of filling up a questionnaire (or a thousand of them).

7. What data?
   As mobile phone data we must a priori understand a set of individual registers with an anonymised/pseudonymised ID, a time attribute, a geospatial attribute, and possibly additional attributes of both sociodemographic and technical nature. This set, especially the geospatial attribute, must be connected to the geographical reference of the national territory. The level of granularity and details of these data sets are part of the negotiation, especially according to legal, privacy, and operational concerns.

8. Define a concrete small research project
   It is important to start with a small business case with concrete goals and concrete

statistical outputs. It must be easily manageable driving both parties to rapid results illustrating both benefits and obstacles to overcome in a next step.

9. Be attentive to legal issues
   It is absolutely fundamental to respect both European and national legal frameworks to build trust, especially, from the citizenship. Show that the statistical public administration is supported by National and European Statistical Acts to access personal microdata. Look for the explicit support of the National Data Protection Authority.

10. Analyse costs
    Data extraction costs (and also possibly data aggregation costs, depending on the business model of collaboration) must be considered as an important piece of analysis of the joint work.

11. Be transparent
    Inform transparently about the scope of the use of mobile phone data and the firm commitment towards confidentiality and privacy protection.

5

# Conclusions for SGA-1

This deliverable closes the activities of the work package for the SGA-1, all of them concentrated upon having access to mobile phone data for official statistics purposes.

We began our activities by designing and administering a questionnaire to take stock of the current situation of access to mobile phone data within the ESS. We provided a set of diverse aspects regarding this access, going from the characteristics of the MNOs over the legal issues to diverse access conditions and data characteristics. We collected responses from nearly the whole ESS to conclude that only a minority of countries has contacted and started some kind of negotiations to access these data and furthermore only a fraction of them has indeed succeeded. In all cases, for the time being, data sets are currently only for research purposes which will allow us to conduct our activities for the SGA-2.

The question to the access to these data is still far from beyond satisfactorily solved for standard production conditions. Apart from the evident methodological and quality concern about what kind of statistics we can produce or streamline with this new information source (to be partially conducted in the SGA-2), having access to these data is an extremely intrincate matter of legal and business issues, not to mention privacy protection or operational/technical models to achieve sharing these data in optimal conditions.

In the direct contact with MNOs (Luxembourg meeting between MNOs and statistical officers), parternships stand clearly as the preferred scenario to solve all these issues. The mobile telecommunication market is noticeably competitive and each MNO shows each own strategy regarding the monetisation of data. Finding the optimal conditions for collaboration in each case requires more contacts and efforts to overcome the different issues in this scenario. The interplay between the national and European levels is and will continue to be an important piece in finding an overall solution in the ESS. As a compromise during the Luxembourg meeting, the ESS will make the effort to continue

to keep these contacts alive and will project further activities jointly between MNOs and the ESS beyond the framework of the present ESSnet.

From the results of the former questionnaire, from the conclusions of the Luxembourg meeting, and mainly from our own experiences in the negotiations with MNOs to access their data, we have compiled a set of guidelines with the goal of aiding ESS partners to conduct their own negotiations. If collaborative partnerships are the preferred choice, NSIs must be ready to show the value of their contribution, mainly through their statistical expertise to deal with inference problems with respect to whole populations and through the enrichment of mobile phone data with complementary official data. So far, official statistics have been built out of questionnaires previously designed and administered with the focus on concrete items. The situation now is completely different. We have large amounts of data and we must be ready to extract the best statistical output of them.

For the SGA-2, once we have several sets of mobile phone data from different countries, we will focus on the methodological, technological, and quality aspects necessary to process them and to produce concrete statistical outputs. Different aspects of data availability (level of granularity, spatial and time coverage, type of events, ...) are definitively limited by the agreements themselves finally reached with the MNOs. This will allow us to explore and assess the diverse possibilities coming out of the different situations.

# Bibliography


5, W. P. (2016). Minutes of the workshop on public – private partnership for mobile phone data for use in official statistics. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/c/c7/WP5_Minutes_22-23_09_2016_Luxembourg.pdf. Luxembourg, 22-23 September, 2016.

ESS (2011). European Statistics Code of Practice (rev. ed.). http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15.

Eurostat (2016). Mobile communication – subscriptions and penetration. https://data.europa.eu/euodp/en/data/dataset/3CwGpDGDQdzHE1ccZkbjw.

Eurostat, NIT, University of Tartu, Statistics Estonia, Positium, IFSTTAT, and Statistics Finland (2014). Feasibility study on the use of mobile poistioning data for tourism statistics. http://ec.europa.eu/eurostat/web/tourism/methodology/projects-and-studies.

Groves, R. (2011). Three eras of survey research. *Public Opinion Quarterly 75*, 861–871.

Khan, W., Y. Xiang, M. Aalsalem, and Q. Arshad (2013). Mobile phone sensing systems: a survey. *IEES Communications Surveys & Tutorials 15*, 402–427.

Laney, D. (2001). 3D Data management: controlling data volume, velocity and variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Mishra, A. (2010). *Cellular technologies for emerging markets: 2G, 3G and beyond*. Wiley.

Normandeau, K. (2013, September). Beyond volume, variety and velocity is the issue of big data veracity. http://insidebigdata.com/2013/09/12/

beyond-volume-variety-velocity-issue-big-data-veracity/. insideBigData.

on Big Data, U. N. G. W. G. (2016). Recommendations for access to data from private organizations for Official Statistics. http://unstats.un.org/unsd/bigdata/conferences/2016/gwg/Item%202%20(i)%20a%20-%20Recommendations%20for%20access%20to%20data%20from%20private%20organizations%20for%20official%20statistics%20Draft%2014%20July%202016.pdf.

Osseiran, A., J. Monserrat, and P. Marsch (2016). *5G mobile and wireless communications technology*. Cambridge University Press.

Positium (2016). Technical documentation for required raw data from mobile network operator for official statistics. ESSnet WP5 internal technical report.

Salgado, D., C. Alexandru, M. Debusschere, F. Dupont, P. Piela, and R. Radini (2016). Current status of access to mobile phone data in the ESS. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5_Deliverable_1.1.pdf. ESSnet on Big Data WP5 Deliverable 1.1.

Sauter, M. (2006). *Communication systems for the mobile information society*. Wiley.

Sauter, M. (2014). *From GSM to LTE – advanced: an introduction to mobile networks and mobile broadband (2nd rev. ed.)*. Wiley.

Work Package 5 (2016). *Minutes of the Workshop On Public – Private Partnership For Mobile Phone Data For Use In Official Statistics*. Luxembourg, 22-23 September, 2016. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/c/c7/WP5_Minutes_22-23_09_2016_Luxembourg.pdf.