ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Workpackage I
Mobile Network Data

Deliverable I.5 (Standards & Metadata)
First proposed standards and metadata for the production of official statistics with mobile network data

Final version, 23 November, 2020

Prepared by:

Roberta Radini (ISTAT, Italy)

- Tiziana Tuoto (ISTAT, Italy)
- Fabrizio de Fausti(ISTAT, Italy)
- Luca Valentino (ISTAT, Italy)
- Raffaella M. Aracri (ISTAT, Italy)
- Sandra Hadam (Destatis, Germany)
- Sandra Barragán (INE, Spain)
- David Salgado (INE, Spain)

Workpackage Leader:

David Salgado (INE, Spain)
david.salgado.fernandez@ine.es
telephone : +34 91 5813151
mobile phone : N/A

Contents

1

General Introduction and Motivation

The availability of a rigorous, understandable and transparent description of the information contained in the Mobile Network data (MND) as well as their structures and the relationship with the statistical concepts is a crucial point for a full understanding of the statistical analyses and experiments based on MND, particularly for official statistical purposes. As well-known, MND are not designed to produce outputs for official statistics; Mobile Network Operators (MNOs) collect and often store MND to track and monitor the operation of the network, to guarantee device connectivity (like the signaling data), to manage the billing of services provided to customers (personal data of the customers as well as data related to the contract, CDRs , DDRs). In addition, the Mobile Network systems that generate these data, being proprietary, have often specific characteristics that might introduce differences among the physical data collected by different MNOs. Other MND specificities are related to the several technologies that have been introduced in telco environment. All these factors highlight the need to share an accurate description of metadata between National Statistical Institutes (NSIs) and MNOs from a conceptual rather than a logical/physical point of view.

Furthermore, even when involving big data sources, like it is the case with MND, the proposed standard statistical production follows the principles of the ESS Reference Methodological Framework (ESS RMF) for MND, comprising:

1. an input phase, which mainly handles data from MNOs. This is also called data layer;
2. a throughput phase where data are processed, transformed and elaborated. This is also called convergence layer;
3. an output phase, with the production of statistical outputs. This is also called statistics layer.

In accordance, it is also useful to distinguish three different types of data and corresponding metadata, as follows:

1. raw source data, as they appear in the big data source, in this case the MNO databases;
2. intermediate statistical data, i.e. the transformed raw data that make statistical processing possible;
3. usual statistical outputs.

It is worthwhile noting that in the Quality Guidelines provided by the WPK Deliverable, ESSnet Big Data pilot 2 ([1] to the Quality Guidelines) the throughput phase has been split into two phases, i.e. sub-phase 1, Deriving Statistical Data from Raw Data of a Big Data Source, and sub-phase 2, Usage of the Derived Statistical Data for the Production of Statistical Output. This subdivision is particularly useful with MND, given that there are some specific operation/transformation/pre-elaboration that are needed to derive statistical data from MND raw data.

In this deliverable we will focus on the conceptual description of the three types of data and in particular the metadata that describe them. Input data are described mainly by providing a glossary for the source data, section GLOSSARY. The glossary uses a descriptive perspective, as we want to avoid being too technical and we privilege a level of details that allow the non-telco-expert readers to understand the informative content of the data source. However, quite often the terminology and acronyms are those used by the standard technical language 3GPP and other specialised technical literature [2], so to create a bridge between the two words, official statisticians and telco experts. This glossary is designed to be useful to define in detail the information requirements of the MND to be used to produce a statistical product and to define a common language for statisticians and telco experts that does not lead to misunderstandings.

The description of the data and metadata related to the first step of the throughput phase is provided in section 3 of this deliverable. Details are provided to clarify why this first step deserve a specific attention, being characterised by a set of operations that are in common to almost all the statistical output that can be derived by MND.

For the description of the other two types of metadata, the former related to the sub-phase 2 of the throughput phase and the latter related to the output phase, we provide some examples in sections 4 and 5. They are most closely related to the production process of the specific usages, so we can't assume to list a complete set of metadata. We limit ourselves to show examples from the use case assumed by this WPI, as purpose of illustration, mainly because the definition in terms of data and metadata of the output phase should help in understanding the acceptable requirements for data and metadata from the input phase and going on.

In this document, we adopt the perspective of the GSIM Referential Metadata Objects [3], where possible. GSIM (Generic Statistical Information Model) is an internationally recognized reference framework for modelling statistical information; in particular, it allows the definition, management and use of data and metadata throughout a statistical production process. This framework, in recent years, has been increasingly adopted by national statistical offices as a conceptual reference model.

2

The modular structure of the Process

This document does not report the flow of processes and data, nor the detail of processes and sub-processes, but the identification of the conceptual model of the data and a first typing of the information content, which we divide into three classes, as specified before: the source data, the throughput data (i.e. intermediate data) and the output data. Recently, some studies have been underway to verify the applicability of the well-known standard process model for statistical production, the GSBPM, into the analysis and production processes of the big data ([4], [5], [6]). To this regard, the project ESSnet Big Data pilot 2, Implementation component, has dedicated the entire WPF (Process and Architecture) to the definition of a new model able to describe the production process with big data sources. Deliverable WPI.6 on Quality will analyze and align the proposal of the model described by WPF for the usage of Mobile Phone data (MPD).

Independently of the process model that we apply, the crucial step for any kind of analysis and production is related to the analysis of the information sources: Data Understanding. This is fundamental both in a cognitive approach to a new source, i.e. in a top-down approach where starting from a cognitive need we try to verify how a new source can meet these needs; and in a similar way it works in an exploratory/mining approach, i.e. in a bottom-up approach, where we need information on the new source to understand what kind of knowledge is there. In both approaches, and in a mixed approach as well, it is necessary to know/understand the data and therefore it is extremely important to document them through the source metadata. These should be used for the selection of the information required to satisfy the knowledge needs, which can be also defined after an analytical and testing phase. In a privacy-by-design approach, only the information strictly required for the project analysis can be selected, therefore the source metadata should be used to determine the transformation of the source data into the data prepared for further processing. At this stage, information generalisation methods can be applied as one of the techniques to minimise privacy risks.

Figure 2.1 shows a light schema of macro processes and metadata typification when using MND. The access of data from MNO systems should be accompanied by the source metadata, for which a basic Glossary of Terms is provided in section GLOSSARY. The raw data selection and archiving phase is followed by a first phase of data preparation that uses source metadata in input to generate intermediate metadata; they are between the initial raw source metadata and the final production metadata. At this point, as explained in the Introduction, we apply the logic proposed in the WPK of split the throughput phase into two. In the first phase of the elaboration, these metadata have the characteristic of being a bridge that brings the raw data closer to statistical information. This metadata represents the transformation from source raw data to statistical data that can be managed with multiple processes that should be controlled and defined by the NSIs, but can be implemented/executed by MNOs on their own premises on trusted analytical systems. Input privacy processes should be introduced at this stage.

This intermediate data and metadata represents the biggest challenge, since it has to be defined jointly by MNOs and NSIs and may involve a classification of concepts, i.e. some general concepts, such as location, might be common to almost all analyses, others specific concepts can be related to the specific use cases. The general concepts will be discussed in section 2 and some examples from the specific concepts related to the methodologies developed in deliverable I.3 of the WPI "A proposed production framework with mobile network data" will be introduced as well.

Finally, production metadata are the traditional metadata of statistical production, which might be also related to new products if the analyses are devoted to new phenomena, not investigated before. In addition, they might represent a deepening or a greater detail of investigation fields already explored in the statistical production of NSIs.

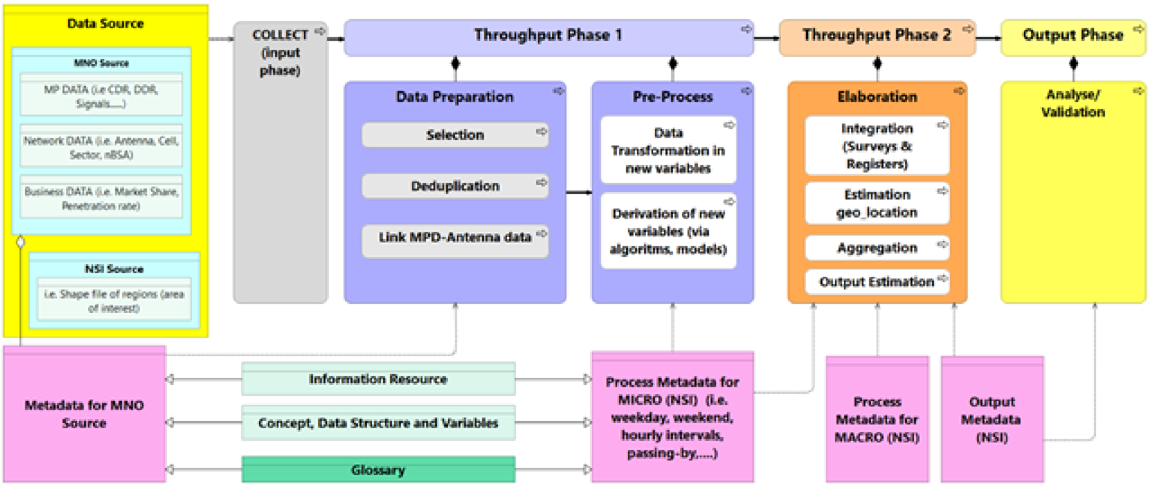
Data Source
MNO Source
MP DATA (i.e CDR, DDR, Signals.....)
Network DATA (i.e. Antenna, Cell, Sector, nBSA)
Business DATA (i.e. Market Share, Penetration rate)
NSI Source
i.e. Shape file of regions (area of interest)
COLLECT (input phase)
Throughput Phase 1
Data Preparation
Selection
Deduplication
Link MPD-Antenna data
Pre-Process
Data Transformation in new variables
Derivation of new variables (via algoritms, models)
Throughput Phase 2
Elaboration
Integration (Surveys & Registers)
Estimation geo_location
Aggregation
Output Estimation
Output Phase
Analyse/ Validation
Metadata for MNO Source
Information Resource
Concept, Data Structure and Variables
Glossary
Process Metadata for MICRO (NSI) (i.e. weekday, weekend, hourly intervals, passing-by,.....)
Process Metadata for MACRO (NSI)
Output Metadata (NSI)

Figure 2.1: Schema of macro processes and typification of metadata

3

The Modular Structure of the Metadata and Standard

The general representation of the analysis and production process of statistical outputs involves the subdivision of data and metadata into: source metadata (Input), metadata (Throughput), and statistical production (Output). These are divided into Micro metadata, if referring to the variables of the individual records, and Macro, if referring to the analysis dimensions of the aggregated data.

The formalization of the metadata follows the GSIM standard. In particular the concept of **Information Resource** [3] is used for modeling the information on the data source, and consists of: the **name** of the source, the **description**, the **owner** (i.e. the MNO for raw source data) and **location** (optional). The scheme of the information resource is shown in the following Table 3.1:

Table 3.1: Schema of the Information Resource for data sources. Taken from GSIM standard [3].

Name	A human-readable identifier for the object
Description	A human-readable description of the object
Owner	Identification of the person, institution or group which owns the information resource
Location	A description of the location where the data resource can be found, it could be a physical address or a logical address (like an URI)

In the case that the data supplies come from multiple providers (MNO) or the process involves the integration of different sources, not just phone data, a table information resource should be defined for each source.

Moreover, the representation of the metadata for each source describes:

- *Data Resource*, i.e. "collections of data that are used by a statistical activity to produce information"[see Data Set in 3];
- *Referential Metadata Resource*, i.e. "collections of structured information that may be used by a statistical activity to produce information"[3]. We formalize the referential metadata resource for MND by means of the glossary of terms, reported in section GLOSSARY.

In this document, we do not address all metadata modeling according to the GSIM model, we rather focalize on some aspects, highlighted with red circles in figure 3.2, which we are sufficient for data and information modeling of MND.

The modeling is carried out on two different layers: a *conceptual* one aimed at modeling the information contained in the data and a *logical-physical* one referred to the data and its structures.

If the logical-physical modelling is purely that of the data and it is shared and defined by the MNO, the conceptual modelling is an operation carried out by the NSI that defines the *units* and the

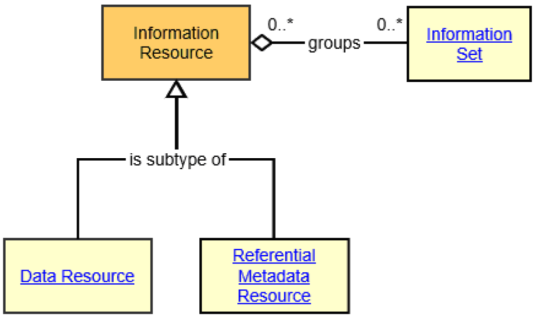

Figure 3.1: GSIM - Information Resource Concept. Taken from GSIM standard **??**.

Figure 3.2: Abstract of GSIM Concept. Taken from GSIM standard [3].

populations of interest. Therefore, the modeling refers to the GSIM concepts of **Data Structure** and **Data Set** for *logical-physical layer*, and to the concepts of **Unit Type**, **Unit**, **Population** and **Variable** for *conceptual layer* (identified in red in figure 3.2).

The *logical-physical* level of the data is described by:

- The **Data Structure** represents the data structure description, it is composed of: *Identifiers*, *Measures* and *Attributes* and can be defined for both Micro and Macro data.
- The **Data Set** is a collection of Data Resource and is structured by a **Data Structure**.

The *conceptual layer* of the data is described by:

- The identification of the analysis units (**Unit Type**) that are represented in the Data Set. A Unit Type is used to describe a class or group of Units based on a single characteristic, but with no specification of time and geography. These contribute to the definition of the population of reference. It is the statistical unit. For example, the calling SIM in the case of a supply /extraction of CDR.
- The individual units that can be extracted from the MNO data supplies, referring to a specific period and an area, represent the **Units**. For example, the Identifier of calling SIMs (single instance of the calling SIM) in the case of a supply /extraction of CDR.
- The **Variable** represents the characteristic of the Unit Type that is to be measured. For example: the Event of Call or Text message, Call start date, Call start time, Call End date, Call End time, etc. in the case of a supply /extraction of CDR.
- The **Population** is made up of a set of units that have homogeneous characteristics in a specific geographical area and a defined time period. For example, a population is made up of SIM subscribers in Rome on October of 2020.

GSIM concepts can also be represented through ontology. Below in figure **??** is an excerpt from the ontological modeling of GSIM concepts in Graphol **??**.

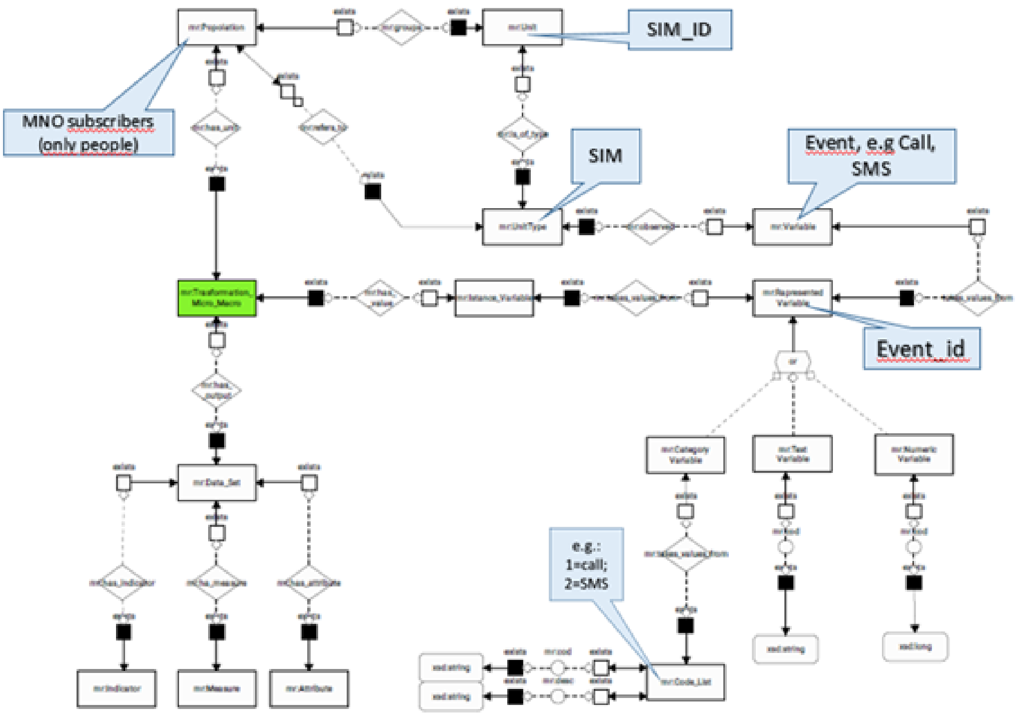

Figure 3.3: The ontology of some GSIM Concepts

In this approach, we follow the idea of using semantics for making data integration, preparation, and governance more powerful. As illustrated in [8], using semantics means conceiving information systems where the semantics of data is explicitly specified and is taken into account for devising all the functionalities of the system. Over the past two decades, this idea has become increasingly crucial for a wide variety of information-processing applications and has received much attention in the Artificial Intelligence, Database, Web, and Data Mining communities [11]. In particular, we concentrate on a specific paradigm, called Ontology-Based Data Management (OBDM), introduced about a decade ago as a new way for modeling and interacting with a collection of data sources ([9], [10]). According to such paradigm, the client of the information system is freed from being aware of how data are structured in concrete resources (databases, software programs, services, etc.), and interacts with the system by expressing its queries and goals in terms of a conceptual representation of the domain of interest, called ontology. More precisely, an OBDM system is an information management system maintained and used by a given organization (or, a community of users), whose architecture has the same structure of a typical data integration system, with the following components: an ontology, a set of data sources, and the mapping between the two.

The ontology is a conceptual and formal description of the domain of interest of the organization, expressed in terms of concepts, concept attributes, and relationships between concepts and logical assertions that formally describe the domain knowledge. The data sources are the repositories accessible by the organization in which the domain data are stored. Mapping is a precise and formal definition of the correspondence between the data contained in the data sources and the elements of ontology. Where element means any concept, attribute or relation. These three levels constitute a sophisticated knowledge representation system that can be managed and reasoned with the help of automated reasoning techniques. Furthermore, the OBDM supports the checking and monitoring of the consistency, accuracy and completeness of data sources.

In general, we can say that the modeling of data and information concepts of MND can be effectively modeled with the GSIM language but this is a language mainly used by statistical institutes while representation through ontologies allows to have a common language with MNOs. They also make the data structures transparent, thus overcoming the heterogeneity of the proprietary information systems of the MNOs.

4

The Source Metadata

We can distinguish the raw source metadata into three types of Data Resources: phone data, network data, and business data. The phone data are generated by the mobile devices directly due to their activities, e.g. calling, receiving a call, sending and receiving text messages, connecting to the internet, as well as indirectly, due to the simple connection to the telco network, even when the mobile devices are inactive, e.g. signaling data.

The network data allow the MNOs to operate the telco networks. This data are related to the characteristics of the telco network, and they are the most technical information referring to: the kind of technologies, the technicalities of the antenna and network.

These terms are familiar for telco experts and are reported in the glossary of terms, due to the fact that a proper understanding of these data is fundamental for using the phone data in the best ways for statistical purposes. Moreover, the knowledge of these terms allows NSI to have a common language with MNO on MPD.

Finally, the business data are related to the business of the MNOs, they represent the number of devices to which the phone data refer; they include info on customer contracts, MNO's market share, and penetration rate at different territorial levels. The **Information Resource** describes the data source used in the analyzes of interest. In our case, the source data are Phone Data, CDR, DDR and/or signaling data, as well as Network Data and Business Data. Moreover, the analyses might involve other data sources, i.e. auxiliary data sources. In this case where the focus is on MPD, those other data can be already in the NSI's data repository or they can be provided by other data owners, e.g. resident population counts at a certain domain, the Land Use, orography for a given territory.

The source data are described with:

- the **Glossary**, according to the schematic of the *Referential Metadata Resource*: Acronym, Lemma and Description;
- the **Data Resource**, classified in 3 types of data: "Phone Data", "Network Data" and "Business Data". Each source data has its data structure.

The Data Structure represents the logical level of the data and it is made up of: *Identifiers*, *Measures* and *Attributes*. A **Data Set** is a collection of data that corresponds to a **Data Structure**.

For a correct use of the data source it is necessary to define the data set according to the components of the Data Structure, and also to associate the description of the content to each variable through the glossary. Once the description of the Data Set has been completed, the information model is defined, and in the second step the *Unit Type* and *Population of interest* are defined.

For example, table 4.1 reports some information from a CDR data set related to the data set descriptors, the data structure and the referential metadata.

Table 4.1: Example of data set descriptors, data structure and referential metadata for CDRs

Data Set Descriptors	Data Structure	Referential Metadata Resource

caller's phone identifier (IMEI) (transformed into ID_SIM);
receiving phone identifier (IMEI) (transformed into ID_SIM);
cell locked by the caller (Cell ID) at the start of the call;
cell locked by the caller (Cell ID) at the end of the call
	Identifier component	IMEI
		ID_SIM
		Cell ID

Call start date;
Call start time;
Call End date;
Call End time
	Attribute component	

Call duration	Measure Component	

It is worthwhile noting that the SIM ID is an identifier in the phone data, while the Cell ID is an identifier in the Network data and it allows us to connect phone activities and the telco network, a crucial step for assessing the location of the phone.

In the case of CDRs we can identify multiple **units of analysis** (Unit Type), for example:

1. the calling SIMs that identifies the active devices;
2. the call events for each SIM;
3. the relationship between the caller and called.

The **Unit Types** characterize a set of Units. This is an example of how the same data set (CDRs) allow analysis on different aspects distributed over time and space.

A. the calling SIMs that represent the "population of active devices";

B. the call event represents telephone traffic;

C. the relations between the caller and called represent the network of telephone contacts.

In particular, in example 1, the Unit Type is "the SIM that makes the call and identifies an active device", while the "Units" are the SIM (IMEI). The set of "Units" analyzed over time and space represents a population, that is, all SIMs active in a certain place and time. This is the typical input data for the density of people analysis in a certain place and time.

With these examples, we want to demonstrate how it is possible to add semantics to the data starting from a set of data by means of a large and in-depth analysis of the information structure, by promoting the extraction of knowledge and allowing to infer new knowledge.

5

The Intermediate Metadata

In this paragraph, we would like to describe the concepts and metadata related to:

- **Data preparation**, i.e. the selection and possible transformations of the data. For example, i.e. the data not to the antenna / sector, but to the Best Server Area (BSA), or generalize the temporal information by transforming the start time of the call from hours/minutes into hours, anonymize the SIM identifier. These processes can be agreed with the MNO and often constitute a constraint for input privacy.

- **Data elaboration and estimation**, i.e. prior/posterior location, estimation density of population by number of SIM.

Intermediate metadata corresponds to those production tasks embedded in the throughput phase of the ESS RMF. These metadata are closely linked to the methodology adopted to process, transform, and prepare data for statistical purposes. We will not provide here mathematical or technical details about the statistical methodology of this phase. We refer the reader to the deliverable WPI.3 "A proposed production framework with mobile network data". However, we briefly describe in generic terms the approach to motivate the Intermediate Metadata and intro duce the context of the terms included in the glossary.

The bottom line of the throughput phase in the ESS RMF aims at detaching the underlying complex technological layer behind the process of generation of MND from the statistical analysis driving us to the final statistical products. MND constitute a rich source of information, specifically about geolocation, Internet traffic, and social interactions. So far, the ESS RMF focuses only on geolocation information. To detach the data layer from the statistics layer, the core idea is to *compute the probability of location of each mobile device at each tile of a given reference grid*. Source data and metadata are used to carry out the computation of these probabilities so that the information coming from this data source is condensed in this set of so-called **location probabilities** for every mobile device anonymously identified. The location probabilities will constitute the basis for any subsequent data processing and modelling exercise, so that source data and metadata are not necessary any more. It is important to clarify and clearly underline that we do not mean that location probabilities, already independent from source data and metadata, can be openly disseminated. They are still highly sensitive data, thus all safeguards regarding privacy and confidentiality must still be applied on them.

Also, to describe the Intermediate Metadata related to the throughput phase we shall follow the same conventions used above for the Source Metadata. In this sense, the structure of the glossary for this production phase is the same. Also, the description of the data sets for the location probabilities runs along similar lines in terms of descriptors, data structure (identifier, attribute, measure) and referential metadata resource (see table 5.1).

Table 5.1: Example of data set descriptors, data structure and referential metadata for location probabilities.

Data Set Descriptors	**Data Structure**	**Referential Metadata Resouce**
caller's phone identifier (IMEI) (transformed into ID_SIM); tile ID	Identifier component	IMEI ID_SIM Tile ID (Reference Grid)
location reference time period	Attribute component	
location probability	Measure Component	

Regarding the Unit Types, what we stated for Source Metadata remains valid, since the throughput phase amounts to computing and assigning measure components (location probabilities and device multiplicity probabilities) for the same units of analysis.

6

The Output Metadata

The Output Metadata are clearly product-oriented, thus intimately related to the statistical domain at stake. From a general perspective, thus, it is impossible to provide a minimal comprehensive list of terms comprising all statistical domains of applications of MND. However, not completely novel terms are needed in this respect, since there already exists a wealth of metadata related to many statistical products obtained with traditional data sources. For example, in tourism statistics rigorous definitions of domestic, inbound, and outbound tourists exist so that regarding MND only a connection between these concepts and the output from the MND-based statistical process needs to be provided.

In our view, this connection must be mostly operational, and only a disruption in the concepts should be introduced if definitively necessary. The focus should be on the algorithmic operationalization of these concepts. In the traditional production setting, concepts and definitions in the metadata system are introduced in the production process mainly through the questionnaire design prior to data collection. Now, data already exists before data acquisition by NSIs and a new problem arises in which those concepts and definitions must be identified through some algorithmic procedure among these existing data.

Let us consider the example of inbound tourism, which can be defined as comprising the activities of non-residents travelling to a given country that is outside their usual environment, and staying there no longer than 12 consecutive months for leisure, business or other purpose. This is a conceptual definition, which can be formalized in terms of GSIM as usual. Regarding MND, we now need to operationalize it, i.e. we need to provide a parameterizable algorithm upon MND producing an identification of inbound tourists in our mobile network data set. Notice that this is intimately linked to the development of the methodology, which is in construction.

In summary, the Output Metadata construction should concentrate on the algorithmic operationalization of traditional statistical concepts and definitions.

Bibliography

[1] https://webgate.ec.europa.eu/fpfis/wikis/pages/viewpage.action?pageId=324045012#QualityGuidelinesfortheAcquisitionandUsageofBigData(DraftinProcess)

[2] https://www.3gpp.org/about-3gpp/about-3gpp

[3] https://statswiki.unece.org/display/clickablegsim/GSIM+on+a+page

[4] Ricciato, F., G. Lanzieri, A. Wirthmann, A., and G. Seynaeve (2020). Towards a methodological framework for estimating present population density from mobile network operator data. Pervasive and Mobile Computing 68, 101263.

[5] Kuonen, D. and L. Bertrand (2019). Production Processes of Official Statistics and Analytics Processes Augmented by Trusted Smart Statistics: Friends or Foes? Statistical Journal of the IAOS 35 (4), 615–622.

[6] Wirth, R. and J. Hipp (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pages 29–39.

[7] Console, M., D. Lembo, V. Santarelli, and D.F. Savo (2014). Graphol: Ontology Representation through Diagrams, in Proc. of Description Logistics 2014, vol. 1193. Vienna, Austria.

[8] Lenzerini, M. (2011). Ontology-based data management, in Proc. of the 20th Int. Conf. on Information and Knowledge Management (CIKM 2011), pages 5-–6.

[9] Calvanese, D., G. de Giacomo, D. Lembo, M. Lenzerini, and R. Rosati (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. Journal of Automated reasoning 39(3), 385–429.

[10] Poggi, A., D. Lembo, D. Calvanese, G. de Giacomo, M. Lenzerini, and R. Rosati, R. (2008). Linking Data to Ontologies. Journan of Data Semantics 10, 133–173.

[11] Noy, N., A.H. Doan, and A. Halevy (2005). Semantic integration. AI magazine 26(1), 7.

GLOSSARY

The glossary aims to allow NSIs to use a common language for MND analysts about MNOs data.

Lemmas selected refer to most common and relevant concepts used by MND analysts.

We considered specialized articles and the experiences of the WPI group to identify keywords and definitions. We used Handbook, Guidelines, Deliverables from WPI, and Official Statistics applications. In some cases, the general definitions come from technical documentation made available by specialized websites such as 3gpp. Wherever possible we avoided excessive technicalities or reworked them. The collected vocabulary therefore integrates and harmonizes experiences and concepts from multiple sources without necessary referring to any specific one.

Lemmas are grouped into three distinct classes by subject area:

- **Phone Data** containing the mobile phone data headwords.
- **Network Data** containing the terms that refer to the mobile network, that is the structure that connects the mobile phones, and related aspects like localization.
- **Business Data** containing customer management and contract management systems data, such as billing data, subscribers information and the MNO market information such as market share and penetration.
- **ESS RMF Data** containing the lemmas of the Reference Methodological Framework.

The glossary is a live tool and continuous updates are expected to come, according to the experiences of each NIS working with MND, in a way that all data analysts adopt a common dictionary with clear definitions of the objects required for any statistical analysis.
Reference:

[1] UN Global Working Group on Big Data for Official Statistics (2017). Handbook on the use of Mobile Phone data for Official Statistics. United Nations.

[2] WP5 of ESSnet on Big Data I. Deliverable WP5.2: Guidelines for the access to mobile phone data within the ESS.

[3] Sauter M. (2011). From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband. Wiley.

[4] Ricciato F. (2019). List of variables in Mobile Network Operator signalling records that are of potential interest for Official Statistics applications, DRAFT version 1.1 (29 October 2019).

[5] WPI of ESSnet of Big Data II. Deliverable WPI.3: A proposed production framework with mobile network data.

Data Resource: Phone Data.

Lemma	**Short or acronym**	**Definition**	**Ref.**
Call Detail Record	CDR	Call detail records (CDRs) capture information on calls made on telephone systems, including who made the call (name and number), who was called (name if available, and number), the date and time the call was made, the duration of the call, and typically dozens of usage and diagnostic information elements (for example, features used and reason for call termination). CDRs are collected on a regular basis for processing into usage, capacity, performance and diagnostic reports. With such information, it is easier to spot exceptions to regular calling patterns, such as out-of-hours calling, international calls, significant variances from previous reporting periods and call destinations that do not reflect normal calling patterns for the enterprise. 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).	
Data Detail Record	DDR	Data Detail Records include internet traffic between the mobile devices and the network which is often referred as IPDR (Internet Protocol Data Records). Mobile devices with internet connection turned off or without the internet capability do not produce DDRs. Because many apps installed on smartphones actively use internet connection to exchange information even when the phone is in the pocket, a huge number of DDRs are stored for such devices.	
Domestic data		Any location events occurring within the network of the specific MNO. These are CDR's or other events where home subscriber (a customer) of the specific MNO is involved. A local subscriber calling another local subscriber in the same MNO network will generate at least two domestic events (one call initiation, one call receiving).	2

Lemma	Short or acronym	Definition	Ref.
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.	
Foreign visited network		This is only relevant for outbound roamers. The home operator should be able to observe the MCC/MNC of the foreign visited network. In some cases, the home operator may be able to learn the position of the outbound roamers at sub-country level (e.g. MSC area or SGSN area). If such information is available, it might be helpful for analysing mobility patterns of outbound roamers at a better degree of spatial detail (E.g., region or city of the visited foreign country).	
Inbound roaming data		Any location event by a foreign MNO subscriber. These data usually represent foreign subscribers using a local roaming service. This data can also include domestic subscribers from another MNO using a roaming service because there is no reception by their own MNO.	

Lemma	Short or acronym	Definition	Ref.
International Mobile Station Equipment Identity	IMEI	Every mobile device is uniquely identifiable with either its IMEI. Identification is used to recognize valid devices and ban devices that are lost or stolen from the network. The IMEI is 15 decimal digits long. It consists of three parts. The first 8 digits, the Type Approval Code, are used to identify the model. The next 6 digits are a unique serial number for this type of device. This is followed by either a 1 digit control code. The IMEI was introduced with GSM. As of GSM the identification of the device and the subscriber are separated. A SIM card is used to identify the subscriber.	
International Mobile Station Equipment Identity and Software Version number	IMEISV	IMEISV is a unique number (15 decimal digits long) used to identify a mobile station (MS). It is composed to IMEI and the last digit of IMEI is replaced by 2 digit that identify software version number.	2
International Mobile Subscriber Identity	IMSI	As long as a subscriber has not changed his SIM card, the IMSI will remain the same. It is internationally standardized unique number to identify a mobile subscriber. The IMSI is defined in ITU-T Recommendation E.212. The IMSI consists of a Mobile Country Code (MCC), a Mobile Network Code (MNC) and a Mobile Station Identification Number (MSIN). When focusing on a single country, the MCC can be dropped out and the combination of the MNC and the MSIN, which is a 9-digit code identifying the subscriber, is usually referred as the National Mobile Subscriber Identity (NMSI).	
IP Detail Record	IPDR	Provides information about Internet Protocol (IP)-based service usage and other activities that can be used by operations support systems (OSSes) and business support systems (BSSes).	
Location area identity/ Tracking Area Identity/ Routing Area Identity		The LAI/TAI/RAI consists of three elements where the first two parts of the code are always the same for the MNO (see added lemmas of LAI and TAI).	2

Lemma	Short or acronym	Definition	Ref.
Mobile Country Code and also Mobile Network Code.	MCC or MNC	MCC is the code of the home country of the inbound roamers. It is a three-digit identification of the country where the LA is located. The MCC is defined by the ITU-T Recommendation E.212. The MNC (a code of a home operator) might help to assess the selectivity bias that is in place due to preferential roaming agreements between MNOs. As for home subscribers, we expect that a single MNO might have customers with different MNC as the result of previous mergers and acquisitions. In this case, MNC information might turn useful to segment different customer groups. For this reason, it might be useful to retain the MNC also for home subscribers.	
Mobile device	MD	Mobile devices are portable computing devices equipped with radio transmitters enabling them to connect to a telecommunication network. Additionally, these devices are increasingly equipped with a number of sensors of different nature (accelerometer, gyroscope, digital compass, GPS radio transmitter, . . .). Depending gyroscope, digital compass, GPS radio transmitter, . . .). Depending on the amount of technology (these sensors) and their size, these devices receive diverse names (mobile phones, feature phones, smartphones, tablets, . . .).	
Mobile phone activity		This individual data includes both the passive signaling and mobile phone activity (calls, messages, etc.)	
Outbound roaming data		Any location event by a local MNO subscriber conducted in another network (usually a foreign MNO roaming service). These data usually represent the local subscribers using mobile phones while travelling in foreign countries.	
Probes (Sonde) (Passive)		Call activities and handover logs; personal features from the operator	
Record type		It 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)	1
Release time		It is the time when seized resources are released again. Release time is an optional field	1

Lemma	Short or acronym	Definition	Ref.
Short Message Service	SMS	It is a paging service to send short messages to mobile telephones. SMS started as a service within the GSM network. Nowadays, a number of other systems for mobile communications offer the same service. A Short Message Service has a length of 1120 bits. This gives the possibility to send and receive text messages of at most 160 characters or data messages of at most 140 bytes. However, it is possible to join a few messages to one larger message. Is an evolution of the Short Message Service (SMS), extending the text content with capabilities to transmit multimedia messages to other mobile users. The messages can be any combination of images, animations, audio (voice, music), video and text. MMS supports most common compression techniques, such as JPEG and GIF for pictures, MPEG-4 for video and MP3, WAV and midi for audio.	2
Signaling data		Signaling data are generally referred to obtaining transmission signals from the radio access network (RAN) directly and storing it to database. They are very voluminous, and can be limited to inbound roaming and domestic data, excluding outbound roaming data. They may contain some parameter values related to the radio channel between the antenna and the mobile terminal that are useful to narrow down the possible location of the latter.	3
Subscriber Identity Module	SIM	The Subscriber Identity Module is a smartcard that is necessary to make use of a mobile phone. The SIM is the key used to identify and authenticate the mobile subscriber. On the SIM is also memory available for personalised data, such as a telephone book and messages. The subscriber is identified with an IMSI, International Mobile Subscriber Identity, and a telephone number. The SIM made possible a clear separation between a mobile phone and a subscriber. The subscriber can make use of any mobile phone under his own account if the SIM card is put in the phone. The use of a SIM can be guarded with a PIN code from 4 to 8 digits. If 4 times the wrong PIN code is typed, the SIM will be blocked. To unblock the SIM the PUK (PIN Unblocking Key) is needed. This is an 8 digits code that is known by the authorised user and given by the service provider. The SIM card is essentially an internal integrated circuit card (ICC) providing diverse functionalities to establish an authenticated secure communication between the mobile device and the network.	2

Lemma	Short or acronym	Definition	Ref.
Subscriber or equipment identity		Receiving/sending parties	2
Transferred Account Procedure	TAP	The transferred account procedure (TAP) includes CDRs and IPDRs that are sent to the PLMN (outgoing data) of the roaming subscribers.	2
Type Approval Code	TAC	This is the prefix of the International Mobile Equipment Identifier (IMEI) that encodes the type of device. While the full IMEI is to be considered personal information, and should not be retained, the TAC code should not be considered sensitive from a privacy perspective. The TAC code might be useful, among other things, to improve the filtering of IoT/M2M devices controller	2
Type of Event	ToE	Type of Event (ToE). We are mostly interest to single out events of type Explicit Detach (ED) from all other types of event, therefore the ToE could be encoded with a single-bit flag. In fact, the ED events are relevant for the interpretation of BA.	2
User pseudonym		This is typically obtained by hashing the International Mobile Subscriber Identifier (IMSI) or part thereof (typically, the part that follows the MCC/MNC prefix).	2
Visitor Location Register	VLR	The Visitor Location Register (VLR) is a database in a mobile communications network associated to a Mobile Switching Centre (MSC). The VLR contains the exact location of all mobile subscribers currently present in the service area of the MSC. This information is necessary to route a call to the right base station. The database entry of the subscriber is deleted when the subscriber leaves.	2

Data Resource: Network Data.

Lemma	Short or acronym	Definition	Ref.
3rd Generation Partnership Project's	3GPP	3rd Generation Partnership Project. The joint standardization partnership responsible for standardizing UMTS (Universal Mobile Telecommunication System) , HSPA (High Speed Packet Access) and LTE (Long Term Evolution)	

Lemma	Short or acronym	Definition	Ref.
Access Point Name	APN	This is a logical identifier, formatted as a character string, that is used in 3G and 4G. It is somewhat related to contractual arrangements and type of subscription: a large company purchasing a stock of multiple mobile subscriptions for its employees will be typically assigned a special APN. Maintaining information about the APN might turn useful for the interpretation of spatio-temporal patterns that are specific to particular groups and/or for the selection of particular user groups. I expect that the MNO will use APN values together with TAC codes, IMSI ranges and IMEI ranges to pre-filter IoT/M2M devices, but such filtering might not be perfect. Maintaining APN information could be useful in detecting residual IoT/M2M devices in the data set. It is unlikely that the MNO will be willing to share cleartext APN values with the Statistical Offices, as this information is highly sensitive from a business point of view. In the best case they will pseudonymise the APN (e.g. by some hashing function, as done with the IMSI) but this is not a problem for statistical applications.	
Answer time		It is the time when a call is answered – the connection was successful. Answer time is a mandatory field for successful calls.	
Authentication Centre	AuC	The Authentication Centre 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.	
Base Station Controller	BSC	The base station provides the control over each BTS as well as the connection to the core network. While the base station is the interface element that connects the mobile devices with the network, the BSC is responsible for the establishment, release and maintenance of all connections of cells that are connected to it.	
Base Transceiver Stations	BTS	Base stations, which are also called Base Transceiver Stations (BTSs), are the most visible network elements of a GSM system. Compared to fixed-line networks, the base stations replace the wired connection to the subscriber with a wireless connection, which is also referred to as the air interface. The base stations are also the most numerous components of a mobile network and comprises diverse elements, apart from the antenna itself, with diverse functionalities (encrypting and decrypting communications, spectrum filtering tools,. . .)	2

Lemma	Short or acronym	Definition	Ref.
Billing Centre	BC	The Billing Centre 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).	
Bounding Area	BA	The BA information is potentially useful when large temporal gaps exist between two consecutive event records for the same mobile device: it provides a constraint of the possible position of the mobile device between two consecutive observations. In order to refer collectively to the entities listed below - Location Area (LA) in 2G (and in the Circuit witched (CS) section of 3G). ▪ Routing Area (RA) in 3G (specifically, the PacketSwitched (PS) section thereof). ▪ Tracking Area (TA) or, if enabled, Tracking AreaList (TAL) in 4G. In other words, depending on the Radio Access Technology (RAT) — 2G, 3G or 4G — the BA will map to LA, RA or TA. As for 3G, a generic RA is always entirely contained within a LA: if both RA and LA are available in the same event, the BA will map to the smallest of the two, i.e., RA	
Cell global identification	CGI	CGI is the unique identifier of a single cell	
Cell Identifier	Cell ID	This information is available for home subscribers and inbound roamers. The Cell ID can be either in the form of a MNO-specific identifier (according to a local cell naming convention) or in the form of the standard Cell Global Identifier (CGI). Both options are acceptable. The cell ID should be interpreted as a pointer to the cell coverage area. This identity is unique only inside the location area	
Cell sites		Collection of transmitters at one specific location	
Cell tower		Collection of transmitters at one specific location that can be referred to as a cell site	

Lemma	Short or acronym	Definition	Ref.
Central Storage Systems		The central storage systems are composed by: ▪ 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.	
Circuit Switched	CS	Circuit switching is a method whereby a dedicated physical path, or circuit, is established and maintained between two nodes or locations for the duration of a connection. Circuit switched networks are often referred to as connection-oriented networks because the dedicated circuit must be estalished first, or “nailed up”, before information can be sent. Telephone networks are typically circuit switched, because voice traffic requires the consistent timing of a single, dedicated physical path to keep a constant delay on the circuit. The plain old telephone system (POTS) is the largest circuit switched network. The original GSM network is also circuit switched. Although GPRS introduced packet switching in the GSM network.	
Core Network	CN	A core network is a telecommunication network’s core part, which offers numerous services to the customers who are interconnected by the access network. Its key function is to direct telephone calls over the public-switched telephone network. In general, this term signifies the highly functional communication facilities that interconnect primary nodes. The core network delivers routes to exchange information among various sub-networks. When it comes to enterprise networks that serve a single organization, the term backbone is often used instead of core network, whereas when used with service providers the term core network is prominent. This term is also known as network core or backbone network.	

Lemma	Short or acronym	Definition	Ref.
Gateway GPRS Support Node	GGSN	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).	
General Packet Radio Service	GPRS	The General Packet Radio Service enhanced the GSM standard to transport data in an efficient manner and enabled wireless devices to access the Internet.	
Global System for Mobile Communications	GSM	The Global System for Mobile Communications is also known as 2G, the predecessor of the 3G network.	
GSM-like mobile telecommunication network		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 antennas 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 (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, . . .).	3
Location area	LA	Adjacent cells, typically 30 or 40, are grouped together into one location area. A group of cells form a location area.	
Location Area Code	LAC	The served area of a cellular radio network is usually divided into location areas. Location areas are comprised of one or several radio cells. Each location area is given a unique number within the network, the Location Area Code (LAC). This code is used as a unique reference for the location of a mobile subscriber. This code is necessary to address the subscriber in the case of an incoming call. The LAC forms part of the Location Area Identifier (LAI) and is broadcasted on the Broadcast Control Channel (BCCH).	

Lemma	Short or acronym	Definition	Ref.
Location Area Code (LAC)/Tracking Area Code (TAC)	LAC/TAC	The Location Area Code (LAC)/Tracking Area Code (TAC) is an identifier of the location area within an MNO's network. This part of the code can be represented using hexadecimal values with a length of two octets. The served area of a cellular radio network is usually divided into location areas. Location areas are comprised of one or several radio cells. Each location area is given an unique number within the network, the Location Area Code (LAC). This code is used as a unique reference for the location of a mobile subscriber. This code is necessary to address the subscriber in the case of an incoming call.	
Location Area Identity	LAI	LAI consists of three elements where the first two parts of the code are always the same for the MNO. For registration in the network, the Mobility Management Entity (MME) has to inform the MSC of the 2G/3G Location Area Identity (LAI) in which the mobile device is currently 'theoretically' located. Since this is only a theoretical value, it has to be computed out of the Tracking Area Identity (TAI), which is the corresponding identifier in LTE. In practice, this creates a dependency between the TAI and the LAI, that is, the location areas that describe a group of base stations in 2G/3G and LTE must be configured in a geographically similar way for the fallback to work later on.	
Location Based System	LBS	The proprietary monitoring system in place at the MNO might be already using the radio parameters (TA, received strength and others) to narrow down the position of the mobile terminal. Generally speaking, if LBS data are available, they should be definitely reported in addition to (not in replacement of) Cell ID and TA.	
Location identifier		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).	
Long Term Evolution	LTE	The Long Term Evolution is also known as 4G.	2
Mobile Positioning (Attive) System		Easy for small samples, usually requires opt-in Accurate positioning, custom frequencies; questionnaire with the respondent possible	

Lemma	Short or acronym	Definition	Ref.
Mobile Subscriber Integrated Services Digital Network number	MSISDN	The MSISDN is the mobile phone number. It is composed of the country code, the national destination code and the subscriber number. As long as subscriber has not changed his/her operator thus also changing the SIM card, the MSISDN will remain the same, due to portability of mobile phone numbers.	
Mobile Switching Centre	MSC	The Mobile Switching Centre (MSC) is a telephone exchange that makes the connection between mobile users within the network, from mobile users to the public switched telephone network and from mobile users to other mobile networks.	
Multimedia Messaging Service	MMS	MMS is a store and forward messaging service. This means that if the recipient phone is not switched on, the message will be stored in the network and sent to the recipient as soon as the phone is switched on. Different protocols can be used as MMS transport mechanism, such as WAP, HTTP or Session Initiation Protocol (SIP). The most common approach used nowadays is WAP.	
N-Best Servers Areas	N-BSA	It is able to compute and report, for each point in space (typically rasterised, e.g. at the granularity of 50 m x 50 m), the predicted signal level of the N strongest radio cells. These maps are called N-Best Servers Areas (N-BSA)	
Network antennas location		The minimum geographical information provided by the MNOs and it corresponds to the coordinate pair of the antenna point	
Network cell		Every cell can be described by a number of attributes such as azimuth, sector angle, shape and size of the coverage area, type of antenna and location.	2
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.	
Network Event		It is information recorded by the telecommunications network that is generated when a device on the network interacts with other elements of the network itself. At each interaction, a signal recorded as an event is generated. Some of these, such as the interaction of the mobile device with the antenna, are of particular interest because they are used to manage the network itself.	
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.	

Lemma	Short or acronym	Definition	Ref.
Operation and Support Subsystem	OSS	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.	
Public Land Mobile Network	PLMN	The Public Land Mobile Network is the entire network of the operator, where each BTS is associated with a geographical cell covering this territory of the PLMN.	
Radio Access Network	RAN	A radio access network is a technology that connects individual devices to other parts of a network through radio connections. It is a major part of modern telecommunications, with 3G and 4G network connections for mobile phones being examples of radio access networks.	
Radio Access Tecnology	RAT	2G, 3G or 4G	1, 4
Routing area	RA	It is the counterpart of LA in packet-switched (PA) networks. The RA is usually a smaller area compared to the LA because using multimedia services requires more frequent paging messages. Reducing the area that needs to be paged helps to lower the number of paging messages that are sent out.	2
Routing Area Code	RAC	Routing Area Code (RAC), which is a one octet long code	1
Routing Area Identity	RAI	For packet-switched networks, Routing Area Identity (RAI) plays the same role as LAIs/TAIs.	1
Seizure time		It is the time when resources are seized to provide service to the subscriber. This field is mandatory only for calls that were unsuccessful	1
Serving GPRS Support Node	SGSN	Serving GPRS Support Node (SGSN) is a main component of the GPRS network, which handles all packet switched data within the network, e.g. the mobility management and authentication of the users. The SGSN performs the same functions as the MSC for voice traffic. The SGSN and the MSC are often co-located.	1

Lemma	Short or acronym	Definition	Ref.
Time Advance	TA	If signalling data are extracted from the Radio Access Network (RAN) they may contain some parameter values related to the radio channel between the antenna and the mobile terminal that are useful to narrow down the possible location of the latter. Timing Advance (TA) that is related to the round trip propagation delay between the antenna and the mobile terminal at the time the message was generated, and therefore is linked to the (quantised value of the) physical distance between antenna and mobile device.	2
Tracking Area	TA	The tracking area is the LTE (Long-Term Evolution) counterpart of the location area and routing area. A tracking area is a set of cells. Tracking areas can be grouped into lists of tracking areas (TA lists), which can be configured on the User Equipment (UE). Tracking area updates are performed periodically or when the UE moves to a tracking area that is not included in its TA list. Operators can allocate different TA lists to different UEs. This can avoid signaling peaks in some conditions: for instance, the UEs of passengers of a train may not perform tracking area updates simultaneously.	2
Tracking Area Identity	TAI	For registration in the network, the Mobility Management Entity (MME) has to inform the MSC of the 2G/3G Location Area Identity (LAI) in which the mobile device is currently "theoretically" located. Since this is only a theoretical value, it has to be computed out of the Tracking Area Identity (TAI), which is the corresponding identifier in LTE. In practice, this creates a dependency between the TAI and the LAI, that is, the location areas that describe a group of base stations in 2G/3G and LTE must be configured in a geographically similar way for the fallback to work later on.	2
Universal Mobile Telecommunication System Network Event	UMTS	UMTS, also known as 3G and the most prevalent mobile communication technologies, follow the 3rd Generation Partnership Project's (3GPP) technical specifications. It is information recorded by the telecommunications network that is generated when a device on the network interacts with other elements of the network itself. At each interaction, a signal recorded as an event is generated. Some of these, such as the interaction of the mobile device with the antenna, are of particular interest because they are used to manage the network itself.	

Data Resource: Business Data.

Lemma	**Short or acronym**	**Definition**	**Ref.**
Contract type		Possible contract types are pre-paid, post-paid SIM, machine-to-machine SIM.	
Customer data	CRM data	Customer databases contain information about users, which can add extra value to CDR or IPDR data – e.g., socio-demographics and place of residence.	
Customer Relationship Management	CRM	CRM is an integrated management information system that is used to schedule, plan and control the sales and pre-sales activities in an organization. CRM systems comprise of hardware, software and networking tools to improve customer tracking and communication.	
Home Location Register	HLR	The Home Location Register is a database from a mobile network in which information from all mobile subscribers is stored. The HLR contains information about the subscriber's identity, his telephone number, the associated services and general information about the location of the subscriber. The exact location of the subscriber is kept in a Visitor Location Register.	
Mobile Network Operator	MNO	A mobile network operator (MNO) is a telecommunications service provider organization that provides wireless voice and data communication for its subscribed mobile users. Mobile network operators are independent communication service providers that own the complete telecom infrastructure for hosting and managing mobile communications between the subscribed mobile users with users in the same and external wireless and wired telecom networks. Mobile network operators are also known as carrier service providers, mobile phone operator and mobile network carriers.	

Lemma	Short or acronym	Definition	Ref.
Socio-demographic attributes		The socio-demographic attributes of the subscribers are usually collected in the CRM system of the MNO and related to the customer's profiling system. The socio-demographic attributes are mostly collected for post-paid private (non-commercial) customers, as there is often no information on the profiles of pre-paid customers or the information is limited (unless they are required to register upon the purchase of the SIM card, and even then it is not always available in the CRM). For commercial contracts it is often not possible to know which specific person is using the phone (and therefore the attributes of the person are unknown). With post-paid private customers, data are also often biased in case of family plans – the sociodemographics of the person who signed the contract (father, mother) are extended to the whole group (meaning the age and gender of the children using the phone are by extension those of the father's).	2, 4
Subscribers' additional attributes		This is relevant for the MNOs store information about their customers in subscriber databases (for domestic and outbound customers). The information that is collected and stored in the Customer Relationship Management (CRM) system is highly MNO-specific typically includes socio-demographic characteristics of the subscriber (owner of the contract) that might be age, gender, preferred language, etc. as well as details on the contract and service such as private or business client, invoice address, average cost of the service or contract type (pre-paid, post-paid SIM, machine-to-machine SIM). This information should be used with great caution, as it is highly sensitive and not always accurate (phone user may differ from the contract holder). The value of these attributes lies in the possibility to add extra dimensions to statistical analysis (e.g., gender), as well as in the option of performing data cleaning processes (e.g., removal of M2M data).	2
Mobile Phone Market Share		The percentage of the total number of contract subscriptions (sales) that is earned by a particular telephone company over a specified period of time. Market share is calculated by taking the company's sales in the period and dividing them by the total sector sales in the same period. This metric is used to give a general idea of the size of a telephone company to its market and competitors.	
Mobile phone penetration rate		Mobile phone penetration rate is often used to mean the number of active mobile phone users per 100 people within a specific population, which is technically not a penetration rate as it does not account for users having multiple mobile phones and hence can exceed 100% due to double counting.	

Data Resource: ESS RMF data.

Lemma	**Short or acronym**	**Definition**	**Ref.**
Conditional Location Probability		Probability of location of a network event conditioned on another network event immediately before.	5
Convergence layer	C-layer	Intermediate layer of the ESS Reference Methodological Framework focused on the transformation and preparation of data for statistical purposes.	5
Data layer	D-layer	First layer of the ESS Reference Methodological Framework focused on the access and preprocessing of raw telco data in preparation for the statistical processing.	5
Deduplicated penetration rate		Penetration rate adjusted by a deduplication factor.	5
Deduplication factor		Numeric factor adjusting for the device multiplicity circumstance.	5
Device duplicity		Circumstance in which a given individual carries exactly two devices during his/her displacement.	5
Device duplicity probability		Probability that a given device is carried together with another one (and only one more) by a given individual during his/her displacement.	5
Device multiplicity		Circumstance in which a given individual carries multiple devices during his/her displacement.	5
Device multiplicity probability		Probability that a given device is carried together with another ones by a given individual during his/her displacement.	5
Dynamical Approach		Approach to compute location probabilities where a displacement pattern of mobile devices is modelled.	5
(HMM) Emission Model		Probability model for the emission probabilities.	5
(HMM) Emission Probability		In a dynamical approach based on a hidden Markov model, probability of observation of network event data conditioned on the state of the mobile device.	5

Lemma	Short or acronym	Definition	Ref.
ESS Reference Methodological Framework	ESS RMF	Production framework based on modularity and evolvability principles to incorporate mobile network data in the production of official statistics. It comprises three layers named data layer, convergence layer, and statistics layer whose main business functions are to access and preprocess raw telco data, to transform and prepare data for statistical purposes, and to build statistical products for different statistical domains, respectively.	5
Event location probability		(See event location likelihood). Geolocation Generic assignment of coarsed spatiotemporal coordinates. E.g. assignment of a city district to a network event.	5
Geoposition		Generic assignment of spatiotemporal coordinates. E.g. assignment of latitude and longitude parameters to a network event.	5
Joint Location Probability		Probability of location of two consecutive network events for a given mobile device. Location Geolocation of a network event to a tile of the reference grid.	5
(Event) Location Likelihood		Probability of network event data conditioned on a given location.	5
Location probability		Probability of the location of a network event.	5
Posterior location probability		Probability of the location of a network event conditioned on network event data.	5
Prior location probability		Probability of the location of a network event not conditioned on network event data.	5
Reference grid		Grid of reference in the ESS Reference Methodological Framework upon which geolocation will be referred to.	5
Region		Unit of territorial division upon which final estimates will be produced.	
Static Approach		Approach to compute location probabilities where no displacement pattern of mobile devices is modelled.	5
Statistical filtering		Procedure to identify individuals of a target population within a given mobile network data set.	5
Statistics layer	S-layer	Upper layer of the ESS Reference Methodological Framework focused on the construction of statistical products for different statistical domains.	5
Telecommunication network		Techonological infrastructure distributed accross a geographical territory providing a telecommunication service, i.e. a communication service while in movement.	5

Lemma	Short or acronym	Definition	Ref.
Tile		Pixel of the reference grid. (HMM) Transition Model Probability model for the transition probabilities.	5
(HMM) Transition Probability		In a dynamical approach based on a hidden Markov model, probability of transition between consecutive states.	5