GOPA
WORLDWIDE CONSULTANTS

Development, implementation and demonstration of a reference processing pipeline for the future production of official statistics based on multiple Mobile Network Operator data (TSS multi-MNO)

Service Contract Number – 2021.0400

Deliverable 2.2: Updated version of technical documentation for scenarios, requirements, use cases and methods, and high-level architecture

Volume I – Detailed scope, requirements and methodological framework

In association with:

Istat
positium
NOMMON
cbs

Development, implementation and demonstration of a reference processing pipeline for the future production of official statistics based on multiple Mobile Network Operator data (TSS multi-MNO)

Service Contract Number – 2021.0400

Deliverable 2.2: Updated version of technical documentation for scenarios, requirements, use cases and methods, and high-level architecture

Volume I – Detailed scope, requirements and methodological framework

Version: final

Date: 18 November 2024

\ ABSTRACT

The Multi-MNO project aims to **develop, implement and demonstrate a proposal for a reference standard processing pipeline for the future production of official statistics in Europe based on MNO data from multiple operators**. If successful, the proposal developed by the project may be endorsed as European Statistical System (ESS) standard by the relevant ESS bodies. The term 'processing pipeline' refers to the combination of a methodological framework and a reference open-source software adhering to such a framework. The processing pipeline developed in this project will cover an initial set of use cases; nonetheless, it will be designed to be general enough to provide the flexibility and growth capability required to cover other future use cases. The pipeline will be demonstrated and evaluated on real data from multiple MNOs in various EU countries.

This report defines the methodological framework proposed by the Multi-MNO project for the processing of multiple MNO data for official statistics. It comprises the project's conceptual framework, the definition of the reference and demonstrator scenarios and the main details of the data processing flow. This document is the first volume of a series of three, which altogether form the updated version of the technical documentation for scenarios, requirements, use cases and methods, and high-level architecture. The two remaining volumes detail the proposed use cases and the data processing methods.

This project deliverable focuses exclusively on the conceptual and methodological aspects. The technical specifications, the detailed architecture and the software design are defined in other project deliverables.

The report is accompanied by a Glossary introducing and defining the concepts the project is based on.

Authors: The report was prepared under the technical coordination of Tiziana Tuoto (Senior Researcher, ISTAT). Further contributions, expertise and research support were provided by: (in alphabetical order) Gabriele Ascari, Roberto Cenciotti, Erika Cerasti, Loredana Di Consiglio, Cristina Faricelli, Ricardo Herranz, Paolo Mattera, Miguel Picornell, Roberta Radini, Giorgia Simeoni, Margus Tiru, Villem Tonnison, Luca Valentino and Ivan Vasilyev. Specific contributions for the present population use case and the geolocation and continuous time segmentation methods (in Volumes II and III) were provided by: Edwin de Jonge, Jan van der Laan, Matthias Offermans and Martijn Tennekes.

DOCUMENT VERSION STATUS AND FUTURE UPDATES:

The document is a work-in-progress updated version of the conceptual and methodological framework. This version addresses the feedback and comments formulated by the project Advisory Board and other groups of stakeholders on a first interim version of the document. Nevertheless, its content may change in the final revision. This document and any future updates will be publicly disseminated on the Multi-MNO project webpage: https://cros.ec.europa.eu/multi-mno-project

Readers are invited to submit comments and corrections or share their views via email to multimno-project@gopa.de

Abbreviations

AB	Advisory Board
BREAL	Big Data Reference Architecture and Layers
CDRs	Call Detail Records
CGI	Cell Global Identity
EC	European Commission
ESS	European Statistical System
EU	European Union
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GSBPM	Generic Statistical Business Process Model
GSIM	Generic Statistical Information Model
ID	identifier
IMSI	International Mobile Subscriber Identifier
IoT	Internet of Things
IPDRs	Internet Protocol Detail Records
M2M	Machine to Machine
MNO	Mobile Network Operator
MS	Member State
NSI	National Statistical Institute
NSS	National Statistical System
ONA	Other National Authority
PET	Privacy Enhancing Technologies
SDC	Statistical Disclosure Control
SDG	Sustainable Development Goals
SIM	Subscriber Identity Module
SIMS	Single Integrated Metadata Structure
TFMNO	ESS Task Force on the Use of MNO data for Official Statistics

Contents

Index of Figures

Index of Tables

1 INTRODUCTION

This report is the first of a set of three separate volumes that form altogether Deliverable D2.2 of the Multi-MNO project.

1.1 THE MULTI-MNO PROJECT: BACKGROUND AND OBJECTIVES

The data collected by mobile network operators (MNOs) can be a valuable source for the production of official statistics in various policy domains where it is crucial to provide reliable and up-to-date population presence and mobility patterns indicators. Spatial planning, transport and mobility, health and environment, economy and tourism are relevant examples. In recent years, a number of National Statistical Institutes (NSIs) within the European Statistical System (ESS) have started to conduct exploratory activities aimed at using MNO data for the development of innovative statistical products. The activities conducted by ESS members have shown that the use of MNO data for official statistics calls for the development of a set of standardised reference methods and tools adhering to the requirements and principles of statistical production, such as quality, transparency, privacy and scientific rigour.[1] Methodological standardisation provides several benefits:

1. **Comparability**: the use of standardised definitions, classifications, and measurement techniques will ensure that statistical products based on MNO data produce consistent and comparable outcomes.
2. **New business models**: standardisation will favour open-source business models from which both private and public organisations will benefit. Contributions from the open-source community will contribute to building reputation and generating new business opportunities for the private companies that specialise in MNO data analytics products.
3. **Accuracy and reliability:** standardisation will improve the accuracy and reliability of statistical products based on MNO data. Uniform methodologies, guidelines and quality standards for data collection, processing, and dissemination will reduce errors, biases, and inconsistencies, enhancing the credibility and trustworthiness of official statistics based on MNO data.
4. **Transparency, privacy and accountability:** standardisation promotes transparency and accountability in the production of official statistics. By adopting open and well documented methodologies statistical agencies abide to the requirement of documenting how data are processed. This will allow scrutiny, validation and reproducibility of statistical methods by independent experts and stakeholders.
5. **Data integration and exchange:** standardisation will facilitate data integration and exchange at national and international levels. When statistical data adhere to common standards, it becomes easier to combine and aggregate data from different sources, enabling the development of indicators across various domains. Standardised data also supports international comparisons and harmonisation efforts, helping countries align their statistical systems with global frameworks.
6. **Policy formulation and evaluation:** standardised official statistics serve as essential inputs for policy formulation, implementation, and evaluation. Standardisation will ensure that policymakers have access to relevant, up-to-date and comparable statistical products based on MNO data, enabling evidence-based decision-making.

---

[1] See position paper prepared by the ESS Task Force on the use of MNO data for Official Statistics: Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System – 2023 edition - Products Statistical reports - Eurostat (europa.eu)

7. **Public understanding:** standardised official statistics enhance public understanding of complex issues. By presenting data in a consistent and accessible manner, statistical agencies enable individuals, communities, and organisations to comprehend and interpret information effectively, promoting informed discussions on societal challenges.

The Multi-MNO project aims to **develop, implement and demonstrate a proposal for a reference standard processing pipeline for the future production of official statistics in Europe based on MNO data from multiple operators**. If successful, the proposal developed by the project may be endorsed as ESS standard by the relevant ESS bodies. The term 'processing pipeline' refers to the combination of a methodological framework and a reference open-source software implementation adhering to such a framework. The processing pipeline developed in this project covers an initial set of use cases; nonetheless, it is designed to provide the modularity, flexibility and growth capability required to cover other future use cases. The pipeline will be demonstrated and evaluated on real data from multiple MNOs in various EU countries.

1.2 SCOPE AND OBJECTIVES OF THE DOCUMENT

This document defines an updated version of the **methodological framework proposed by the Multi-MNO project** for the processing of MNO data for official statistics, which comprises the project's conceptual framework, the reference scenario and the high-level architecture of the data processing flow. The document is complemented by two other volumes of the same deliverable, which detail the use cases and the data processing methods. It is important to highlight that the deliverable (i.e. all three volumes) focuses on conceptual and methodological aspects; the software's technical specifications, architecture and design are defined in other project deliverables.

1.3 DOCUMENT STRUCTURE

In addition to this introductory section, the remainder of this document is organised as follows:

- **Chapter 2 'High-level requirements'** defines the set of general principles that will guide the development of the processing pipeline.
- **Chapter 3 'Reference scenario'** defines the set of assumptions on which the processing pipeline will be based with regard to aspects such as the European regulatory framework, the data available from MNOs, etc. These assumptions do not necessarily correspond to the situation at the time of writing the present document, but they define a future 'reference scenario' for when the processing pipeline is expected to be used for statistical production based on MNO data. The situation corresponding to the one currently in place, and under which the pipeline will be demonstrated during the project, is called the 'demonstrator scenario'. The differences between the 'reference scenario' and the 'demonstrator scenario' are highlighted in the document.
- **Chapter 4 'High-level data processing flow'** provides an overall view of the proposed data processing flow, stating its fundamental design principles. This includes the required input data, the functional modules and their intermediate results (i.e. inputs and outputs of these functional modules), the final outputs, and the data quality checks that will be implemented to ensure that the statistical outputs meet the level of quality required by official statistics. The proposed data processing flow is intended to be general enough to cover the production of any type of population presence or mobility indicators so that different use cases will be addressed by a specific configuration or parameterisation of the proposed flow.
- **Annex 1 'Overview of project tasks'**
- **Annex 2 'Concepts and definitions'** compiles a consolidated list/glossary of the terms that define the project's conceptual framework. Whenever possible, the project builds on existing, widely accepted concepts and definitions in the domain of official statistics; nonetheless, we also introduce a number of new concepts and definitions whenever existing ones are not satisfactory or not yet adequately consolidated to deal with the processing of MNO data. This section is crucial to build a common language between industry and official statistics. It is conceived as a living glossary that can be enriched each time

the project encounters the need to define and clarify the meaning of new or ambiguous concepts. Therefore, the glossary is a standalone output of the project itself and not exclusively linked to the current deliverable, being relevant to all other project deliverables.

2 HIGH-LEVEL REQUIREMENTS

The generality of the methodological framework of the pipeline is crucial to accommodate a wide range of use cases[2] related to the production of population presence and mobility indicators. This chapter discusses the key considerations to ensure the framework's flexibility, adaptability, and applicability to diverse scenarios. These considerations are the following:

1. **Methodological soundness**: to establish a solid foundation, the framework will incorporate state-of-the-art approaches in the field of MNO data processing. By leveraging cutting-edge techniques, the framework can benefit from the latest advancements and ensure its relevance in the evolving data analysis landscape.

2. **Integration of previous findings**: the framework should take into account the findings of previous Eurostat and ESS projects. Building upon existing knowledge and experiences can avoid redundant efforts and incorporate proven methodologies for improved efficiency and accuracy. Previous work and results are incorporated into the methodological framework to the extent that they meet our requirements. However, they do not represent a constraint, and we will explore and propose different solutions to overcome current limitations.

3. **Stakeholder consultation**: in order to ensure the reliability and credibility of the proposed methods, they will undergo a consultation process by the project's Advisory Board (AB), composed of a group of external, independent experts who will provide feedback on the feasibility, effectiveness and appropriateness of the proposed pipeline for the intended use cases. Along with the AB, the consultation process will regularly involve domain experts from the ESS (e.g. members of the Working Group and Task Force on MNO data for Official Statistics, Tourism, Quality, etc.). Finally, potential consultation events might be organised, for instance, through presentations to selected fora (such as the GSM Association or other selected consortia and conferences).

4. **Evolvability**: the framework must be able to incorporate future methodological improvements in the domain of MNO data processing without requiring a full, in-depth redesign. As new techniques and approaches emerge, the framework will be adaptable and flexible enough to integrate these advancements seamlessly, thereby keeping it up-to-date and relevant over time. Evolvability is most needed in response to the rapid changes in mobile technologies and business models, potentially impacting the MNO data generation systems. In addition, mobile services and usage behaviour evolve over time, and improved procedures and advanced methods are available that can better represent the phenomena of interest compared to the methodology currently used. Furthermore, statistical needs change with emerging demands for new topics and cases. The methodological framework for processing MNO data needs to be designed as an evolving standard to facilitate continuous adaptation and be ready for advancements in the methodologies, new requirements and improved techniques.

5. **Methodological challenges and recommendations**: in cases where mature solutions are lacking in the state-of-the-art, the framework will accurately describe methodological issues, identify their potential impact on the data processing and provide recommendations for future research. By addressing and highlighting these challenges, the framework can stimulate further exploration and innovation in the field.

6. **Modularity**: the data processing flow within the framework is designed with a modular approach. Each functional unit is distinct and clearly specified, including input data objects, output data objects, data

---

[2] The use cases are detailed in Volume II of this deliverable.

processing methods and quality controls. This modularity allows for easy improvement of specific functional units without extensive dependencies on other units, enhancing the flexibility and scalability of the framework.

7. **Use case specific descriptions of the methods using the pipeline**: to cater to different use cases beyond those covered by the project, the framework accommodates the implementation of use case-specific methods. These specialised methods can be implemented as alternative branches in the data processing flow or as alternative configurations/parameterisations of the relevant functional modules. This flexibility allows for easy extension and customisation of the framework for specific requirements.

8. **Consistency**: while different use cases may require tailoring of certain functional modules, the design of the data processing flow will strive to maximise consistency across different statistical products generated by different paths or configurations. This consistency ensures coherence and comparability between various outputs.

9. **Quality assurance**: the framework will incorporate mechanisms aimed at ensuring that the quality of the resulting statistical products is in line with official statistics principles and practices. Quality controls are implemented at different stages of the data processing flow, encompassing input data, intermediate results and final outputs. These controls are designed to facilitate the auditing of each execution of the data processing flow, combining automatic checks and human quality control to enhance the reliability and accuracy of the produced indicators. For example, the data processing flow may generate certain warnings that trigger the need for human review of certain data objects.

10. **Explainability**: is defined as a desirable quality for the data processing flow, particularly when the ground truth for certain statistical populations and indicators is unknown. To maximise explainability, the framework will avoid using 'black-box' methods and favour transparent and interpretable approaches based on explicit rules. This promotes transparency and allows for a better understanding of the underlying methodologies and their implications.

11. **Adherence to standards**: while standards such as the [Generic Statistical Business Process Model](#) (GSBPM), [Generic Statistical Information Model](#) (GSIM), and Big data REference Architecture and Layers (BREAL) can provide valuable guidance, the framework should not be constrained by adherence to any particular standard. Instead, it will strive to leverage relevant standards where applicable while maintaining the flexibility to adapt and evolve as needed.

12. **Openness and reproducibility**: to ensure transparency and facilitate collaboration, the methodological framework will avoid the use of proprietary solutions. All definitions, concepts and methods developed or used by the project will be open and thoroughly documented, ensuring clarity, detail and unambiguousness. This documentation enables methodological reproducibility, allowing others to reproduce and audit the methods independently. Moreover, clear and detailed documentation of the methodology is required to allow the correct interpretation of the results and to stimulate the improvement of methods and procedures.

13. **Multi-MNO orientation**: the standard will be designed to allow information from multiple MNOs to be combined within and across countries. This requirement brings a series of benefits by ensuring a better representativeness of the total population (i.e. reducing the risk of population coverage bias in the final statistics), improving the robustness of the final statistics to anomalies, glitches and interruptions of data availability caused by technical flaws of individual MNOs, and ensuring equal treatment of competing MNOs. The latter prevents the introduction of asymmetries and differences in treatment, reducing the risk of interfering with competition dynamics between MNOs. This approach also ensures that no single MNO holds undue influence, thereby enhancing the NSI's independence from any particular MNO or brand. Additionally, the combination of multiple MNO data offers an extra layer of protection for business sensitive information of a single MNO. For an overview of the solid reasons behind a multi-MNO orientation, we refer to the position paper by the ESS Task Force on the use of MNO data for Official Statistics ([ESS Task Force on MNO data for Official Statistics, 2023](#)).

14. **Privacy protection**: given the sensitivity of data involved in MNO data processing, privacy protection is of utmost importance in this project, as well as during tests and development. The application of the

methodological framework developed in this project will be complemented by appropriate organisational and technical measures to safeguard privacy and business sensitivity. Technical privacy protection measures include pseudonymisation of raw data, secure processing of disaggregated data within a secure computation environment deployed at MNO premises, and the application of statistical disclosure control (SDC) methods to ensure full anonymisation of any data extracted from the MNO's secure environment. The development of new SDC methods is not a requirement of the project.

In summary, the general character of the methodological framework will ensure its adaptability, robustness and applicability to a wide range of use cases related to population presence and mobility indicators. By incorporating state-of-the-art approaches, leveraging previous findings, and adhering to principles of modularity, consistency, quality assurance and privacy protection, the framework will provide a solid foundation for the production of reliable and valuable statistical products. Additionally, emphasising openness, explainability and evolvability will further enhance the framework's usability, transparency and longevity.

3 REFERENCE SCENARIO

The purpose of this chapter is to specify the **reference scenario** for which the data processing pipeline will be developed, i.e. the set of assumptions about the context in which the proposed pipeline will be used by the NSIs. It should be noted that this reference scenario does not necessarily correspond to the situation during the development of the pipeline, nor to the conditions under which the pipeline will be demonstrated and evaluated within the present project. Therefore, we are introducing the so-called **demonstrator scenario**, which is the set of conditions in place during the project in which the proposed data processing pipeline will be demonstrated. The demonstrator scenario is mainly determined by the current privacy protection legislation and the concrete data access and elaboration agreements stipulated with the MNOs participating in this project. If there is a gap between the hypothetical reference scenario and the actual demonstrator scenario in which the methodological framework is finally tested in practice within the project, the difference between them, as well as the impact of such a gap on the project conclusions, will be clearly stated and documented.

3.1 AVAILABILITY OF DATA FROM MULTIPLE MNOS

The reference scenario is based on the assumption that conditions for access to MNO data will be favourable for the re-use of data for statistical purposes by statistical authorities. The re-use of MNO data by NSIs would be based on a sustainable partnership model[3], respecting the strongest possible technical and organisational data confidentiality measures to protect individual privacy and business sensitive information. In any case, microdata – including raw MNO event data and any element of individual data, e.g. 'home location', movement lists, etc. – will not leave the protected computation environment at the MNO and will be processed locally at the MNO. As far as aggregate data are concerned, in the reference scenario, these will be accessed by the NSI without any additional protection measures (e.g. SDC). We acknowledge the fact that aggregate data may still carry a small residual risk of personal re-identification if they are not 'anonymised', and they may still be business sensitive if absolute numbers are provided. However, in the reference scenario, we assume that an appropriate legal basis is in place to allow the NSI to receive such data from all major MNOs in the respective country, to combine them and then apply protection measures (i.e. SDC) to the total aggregate across all MNOs before publication.

For the demonstration of our pipeline in the context of the current project (i.e. in the demonstrator scenario), we might have to accept that the aggregate data will be accessed by the NSI after some kind of SDC (e.g. a k-anonymity filter that removes counts <k) and, possibly, but not necessarily, some form of grossing-up to compensate for the business sensitiveness of the absolute numbers. The value of k and whether or not some form of grossing-up/rescaling is really needed will be discussed with the MNOs involved in the testing and demonstration phase of the project. In any case, these aspects have a minor impact on the overall design and implementation of the methodology.

It is worth noting that both the reference and the demonstrator scenarios assume that the combination of data from multiple MNOs, as well as the integration of MNO and other NSI data, is done exclusively at the aggregate data level, rather than at the micro-data level. The latter case is not considered in this project. An exception here is represented by the statistics based on roaming data, where some form of micro-level integration of inbound roamers from multi-MNO data might be unavoidable. This will be developed as the project evolves. For the current

---

[3] While the elaboration of this partnership model is not part of the project, this could demonstrate the feasibility of aiming towards such a public-private collaboration.

version of this document, the above considerations and distinctions between the reference and demonstrator scenarios apply to national subscribers and outbound roamers.

To tackle both reference and demonstrator scenarios, the pipeline design will be modular: modules can be placed in any infrastructure location and set up in a way that would satisfy both scenarios.

The fact that the consortium includes two MNOs from a single country, namely Orange Spain and Vodafone Spain, will allow us to test the flexibility of the framework in this respect as well.

3.2 CHARACTERISTICS OF MNO DATA

We assume that in the reference scenario, MNOs will always provide two types of data:

\ EVENT DATA

Event data consists of different types of events that provide the geolocation data of the MNO subscribers.

- This will include data from different generations of network technologies (2G, 3G, 4G, 5G).
- We assume that we are dealing with signalling data, not just Call Detailed Records (CDRs). However, the methods will be able to process CDR data in the unfortunate event that signalling data are not available.
- Inbound and outbound roaming records will be available.
- Before making the data available for processing, the MNO will pseudonymise the identifier of the user by applying a hash function. We assume that the pseudonymised identifier of a SIM will remain available for the project's elaborations within the MNO premises for an agreed period of time, i.e. that the MNO will not periodically change the hash function used for pseudonymisation or, if it does periodically change the pseudonymisation, that this change will not affect the processing pipeline.
- In most cases, these data will identify the cell to which the device is connected at the time of the event; it is also possible that some MNOs will provide a more precise position estimate for these events. The pipeline processing will accommodate both cases.
- We assume that the MNOs will only provide event data corresponding to mobile phones, i.e. that they filter out events corresponding to secondary devices (e.g. tablets), M2M/IoT devices, etc.

\ NETWORK TOPOLOGY DATA AND/OR CELL COVERAGE PLAN

- If the MNOs can provide the cell coverage plan, this is the simplest option. In this case, the preferred option is for them to provide the n-best cell coverage areas (with n >= 6), rather than the single best coverage area.
- If they cannot provide the cell coverage plan, the MNOs will provide the necessary information on the network topology required to estimate cell coverage (antenna position, azimuth, type of cell, etc.).
- We assume that this information is updated on a daily basis so that it always reflects the status of the network.

Additional information potentially available from the MNO that our pipeline will be able to exploit to improve the quality of the results includes:

- Data from MNO-operated apps may be added by some MNOs, and they will be supported by the pipeline, even if it is not considered a mandatory input. Data from other app providers are not considered an input for the pipeline.

- Information on some specific groups of contracts/devices, e.g. foreign visitors[4] (if not already identifiable via the first digits of the pseudonymised device_id), special client contracts (e.g. taxi drivers) and users coming from other operators.

Finally, it is worth remarking that customer profile data, which provides information on the sociodemographic profile of the MNO subscribers (e.g. age, gender), are not considered input data for this project, mainly because of their uncertain quality and additional complications related to privacy protection.

3.3 DATA PROCESSING ENVIRONMENT

It is assumed that, due to MNOs' data security policies and/or personal data protection regulations, **individual (non-aggregated) data will always remain within the MNO infrastructure and will never be exported to the NSI**. The traditional approach of delivering raw data directly to the NSI will be replaced by a model that involves data processing and aggregation at the MNO level. As clarified above, when distinguishing between the reference and demonstrator scenarios, in the latter, only anonymous aggregate data will leave the MNO premises and only after the application of SDC measures. In the reference scenario, the NSIs will have access to aggregate data but without requiring full anonymisation, assuming that legislation allows this option. Processes that require the fusion of the pseudonymised MNO event data with other data sources at the individual level will therefore take place within the MNO processing facilities or in a secure processing environment based on Privacy Enhancing Technologies (PET).

These choices encompass legal, technical and privacy considerations:

1. By keeping the data anonymised and aggregated at the MNO level, the process aligns with specific regulations in place, ensuring compliance with local telecom and personal data protection laws that safeguard privacy and data protection.
2. From a technical standpoint, maintaining the data on-site at the MNOs is deemed safer. By leveraging the existing IT facilities and infrastructure at the MNO level, risks associated with data transfer and storage are minimised. This decentralised approach provides an added layer of security, reducing the vulnerability of the data during transit.
3. Efficiency also plays a crucial role: the MNOs are equipped with the necessary infrastructure, systems and know-how to handle and process the data effectively, reducing the burden on the NSI or other entities.
4. Lastly, this approach acknowledges the MNOs' responsibility for the privacy of their subscribers: by allowing the MNOs to supervise and manage their subscribers' data, they can maintain a higher level of control and accountability. This ensures that privacy safeguards are maintained throughout the process.

Since this project is not dedicated to investigating methods for integrating MNO data with other non-MNO data sources, the integration will be limited to a few basic cases to demonstrate this integration module in the pipeline.

---

[4] People coming from abroad are 'foreign visitors', while people carrying a SIM issued by a foreign MNO are international 'roamers'. However, 'foreign visitors' do not always map to an international 'roamer' and vice versa.

4 HIGH-LEVEL DESCRIPTION OF DATA PROCESSING FLOW

The reference processing pipeline developed by the Multi-MNO project will be described at four levels:

1. **Workflow level**: provides an overview of the entire process, describing the proposed high-level architecture, the main data objects, and the functional modules. The sequential flow of operations, decision points, and interactions between the different components of the pipeline are represented in **FIGURE 2**. An overall vision of the process flow is provided at this level.
2. **Semantic level:** focuses on defining each pipeline object in detail, specifying the information content of each data object and related sub-objects, and the logic and methods of each function module and related sub-modules.
3. **Syntactic level**: provides a distinctive syntactic definition of each sub-object. This level focuses on specifying the reference format adopted for the objects within the pipeline.
4. **Implementation level**: corresponds to the software design and the code itself, as well as the associated documentation, including comments and explanations within the code. This is the final level of description, detailing how data is stored and processed, and which data structures are adopted for the data.

The present document focuses particularly on the first two levels, i.e. the pipeline level and the semantic level, since they are at the core of the pipeline standardisation purpose of this project. Besides, the syntactic level is also partially subject to standardisation. Therefore, the syntactic level is the object of this document only as far as standardisation is concerned. Apart from the standardisation aspects, the interaction and overlap between the syntactic level and the implementation level are worked out collaboratively between Task 2 and Task 4, and developed in project's Task 4 (software development)[5]. Similarly, some standardisation decisions at the syntactic level that may facilitate the implementation level are the objective of the Task 4 deliverables.

It is worth noting that this project aims to propose standard solutions for the first three levels of the MNO data processing flow, as well as an implementation of these (the 4th level), where the implementation level is not part of the standard itself.

The remainder of this chapter deals with the description at the pipeline level, while the semantic description of the data objects and the methods are addressed in Volume III – Methods and data object (part of this deliverable).

4.1 FUNDAMENTAL DESIGN PRINCIPLES

The proposed pipeline is designed according to the following fundamental design principles:

1. **Standardisation of a common reference grid across all MNO**

A common reference grid (e.g. the INSPIRE geographical grid system) will be used to represent the spatial dimension throughout the pipeline. All intermediate spatial data will be represented according to the chosen common standard grid. This allows data from different MNOs with overlapping service areas to be combined and ensures comparability, across MNOs and over time.

---

[5] For an overview of the project tasks, please refer to Annex 1 – Overview of project tasks.

Note 1: This principle does not preclude that, starting from intermediate data represented on the standard grid, final statistics are aggregated (or projected) to other types of grids (e.g. administrative units). If this is necessary, the module(s) down-streaming from the pipeline will perform the transformation/projection from the standard grid to the output grid. The detailed method for transformation/projection is part of the pipeline and is subject to standardisation. The description of this method is provided in Volume III – Methods and data object – Chapter 18.

Note 2: This principle does not necessarily require that all MNOs provide topology data in the standard format. While this is the preferred option, input data represented in different ways will be transformed/projected to the standardised grid. The detailed method for transformation/projection is part of the pipeline and is subject to standardisation.

2. Multiscale longitudinal analysis

For each individual mobile device, the data processing flow performs an information reduction that is logically organised into three timescales: short-term (one day), mid-term (one quarter or one month) and long-term (one year). The first functional module takes the granular event data (*aka* nano-data[6]) in as input in daily batches, and produces daily summaries based on a set of open-source heuristics, which are stored in the corresponding daily data object. The latter serves as an input source for the second functional module at the next stage, which processes batches of daily summaries (over one month or one quarter) and produces mid-term (monthly or quarterly) output summaries, which are also based on some open-source heuristic logic. Finally, the third functional module analyses sequences of multiple mid-term summaries and determines long-term data labels for the individual mobile device at hand (e.g. place of residence, set of places constituting the usual environment, etc.).

For each timescale (i.e. short-, mid-, long-term), the corresponding 'summaries' are constituted by groups of data elements (variables, labels, etc.) with semantics that are specific to that particular timescale (e.g. the prevalent place of stay in one day, the prevalent place of stay in one month/quarter, etc.). In addition to the statistical methodological motivations behind the possibility of a multiscale longitudinal analysis, this approach also allows the pipeline to differentiate the data storage according to the data minimisation and storage minimisation principles of the GDPR. In this way, the raw data, which is highly exposed to the risk of re-identification, are stored for a shorter period of time, while increasingly aggregated data, for which the risk of re-identification is lower, can be stored for longer periods.

3. Input data accessibility from any point of the pipeline

Any data element of any data object can be read without restriction by any functional module in the pipeline, and can therefore be used as input. This principle is defined as 'input data accessibility'. There are, however, restrictions in terms of what functional module(s) are allowed to determine (i.e. write or update) the value of each data element (see the next 'bottom-up one-way processing' principle below) and when such determination could take place (i.e. the timing of write or update operations).

4. Bottom-up one-way processing

The processing workflow follows the principle of bottom-up one-way processing, from lower to higher timescales, in the sense that the content of the generic data element at one timescale is determined (i.e. written or updated) by the functional module based solely on the data elements (for the same individual mobile device) from the lower preceding timescale. The statistical argument for this principle is the avoidance of potential bias generated by a downward process where daily, near real-time events are (re)written on the basis of the outcomes of mid- and long- term summaries that can only be calculated afterwards, at the end of the periods.

---

[6] Ricciato, F., Wirthmann, A., & Hahn, M. (2020). Trusted Smart Statistics: How new data will change official statistics. Data & Policy, 2, E7. doi:10.1017/dap.2020.7

Note: *This principle does not exclude that some other processing module(s), other than the three described above, may use (read) information elements from higher and lower timescales and combine them together to produce new aggregate intermediate data, in line with the 'input data accessibility' principle defined above.*

5. Separate integration along different dimensions

The overall workflow involves operations along three dimensions, namely: time, space and units (devices). It is recognised that the data reduction process performed by the methodological pipeline entails some forms of 'integration' across each of these dimensions. Whenever possible, for the sake of modularity and clarity (as well as for evolvability), the methodological design shall seek to clearly separate the integration functions along different dimensions into different functional modules. Ideally, each functional module integrates along only one dimension. When integration along two (or three) distinct dimensions has to be performed, the decomposition of the corresponding algorithm into the cascade of two distinct functional modules, each acting along one dimension only, is encouraged. This principle facilitates the definition and interpretation of the methods, as well as their evolvability and improvement.

6. Balancing flexibility and parsimony

The processing flow should always support the possibility of executing some parallel alternative methods in distinct functional sub-modules (each providing output to a distinct data sub-object) for critical operations without forcing the choice of a single one-size-fits-all approach. This is particularly critical for functional modules embedding heuristic methods, where different variants may yield different profiles of strengths and weaknesses that suit different use cases or groups of use cases, whereas a single solution would have to compromise and in extreme cases result in a poor 'average' fit. This applies in particular (but not exclusively) to functional modules embedding heuristic methods. On the other hand, this principle of flexibility ('not just one option') should be balanced by a parsimonious approach that strives to limit the number of supported variants ('not too many') and cluster groups of use-cases with similar profiles in order to reuse the same method.

7. Soft classification and rigorous uncertainty assessment

Whenever a decision is to be made based on the observed data (e.g. selection of one class or label from a pre-defined set), but the available information is not sufficient to make a choice with a sufficient degree of confidence, the pipeline shall provide the option to omit the decision. For instance, if the decision concerns the selection of one class (or label, tag) from a set of pre-defined options (classification, labelling, tagging), the class/label/tag 'unknown' should always be included in the set of allowed options. In this way, uncertainty is not misleadingly hidden by (likely wrong) 'hard' decisions. This has consequences for both the design of functional blocks and data objects.

The application of these principles will be fully explained and remarked in the illustration of the pipeline in the following subsections.

4.2 ELEMENTS AND LEGEND OF THE PIPELINE DIAGRAM

The pipeline is presented in the form of a process flow diagram, which includes different types of components, each with a specific visual representation:

- **Arrows** represent data flows, including both input data and data generated by the functional blocks of the pipeline.
- **Cylinders** represent (sets of) data objects, used to store information received from input flows or generated by functions. Colours are used to distinguish different kinds of data objects: red is used for input data, yellow to identify data generated by pipeline functions, and green for the final output.
- **Rectangles** are (sets of) functions. These functional modules represent actions that are performed on data objects. The result of the operation is generally a new data object; however, there are also cases where the

output of a function is a direct input for a subsequent function and it is not relevant for other components of the pipeline.

- When a function populates a data object, a *vertical arrow* is used to define its relation.
- When a function performs data demultiplexing (i.e. a separated flow is generated for each mobile device), a set of *parallel arrows* is used to represent the operation.

4.3 INPUT DATA OBJECTS

The input data includes three different data categories, which in turn contain different types of data objects (see **FIGURE 1** below).
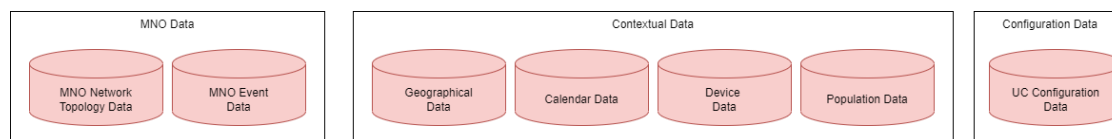

Figure 1: A representation of the input data objects for the project

The different categories are identified on the basis of their role and their usages in the pipeline. In detail, these are:

\ **MNO DATA** including:

1. **MNO network topology data**: these include the cell_id and the coverage areas of the network cells or, alternatively, data that allows the estimation of such coverage areas, such as the location of the antennas, their orientation, characteristics (e.g., the technology: 3G, 4G, 5G, etc.), emission power, etc.
2. **MNO event data**: they result from transactions/events between the mobile device and the network infrastructure. Basically, they report the event as a triplet $<k,i,t>$ for device k connecting with the cell i at the time t. They should at least include CDR/IPDR data and/or signalling data.

The pipeline is specifically designed to analyse these MNO data, both event and network topology data. In principle, they are continuously ingested by the pipeline and therefore represent a continuous flow of data. Unlike the other data objects, they only enter the pipeline at the very beginning, while the other data objects can potentially be accessed by the pipeline at any step as required.

\ **CONTEXTUAL DATA** including:

1. **Geographical data**: this includes grids, administrative boundaries, topography, street networks, land use, hydrography, and coordinates of points of interest. The geographical reference data can be enriched with MNO network topology data; for example, cells dedicated to hospitals, stations, airports, etc.
2. **Calendar data**: tagged time information, e.g. working day, weekend, bank holidays, special events, etc. The calendar data can be enriched with MNO network topology data to help identify particularly relevant events (e.g. temporary cell placement or cell re-purposing linked to a sports event, a cultural event, the launch of a new offer, etc.). Herewith we collect all the info with a strong 'calendar effect', regardless of whether they have been provided by MNO or are publicly available.
3. **Device data**: if available from MNOs, any kind of potentially useful data about the users of the mobile devices, e.g. foreign visitors (if not already identifiable via the first digits of the pseudonymised device_id), special client users (taxi drivers, etc.), users coming from another operator, etc.
4. **Population data**: encompassing census records, tourism statistics and related information. It includes data collected through national or regional censuses or other sources. These data are aggregated to a geospatial level, which can still be very detailed, e.g. a 100-metre grid describing the percentage of females or the number of residents based on the census.

Clearly, the information that binds MNO network topology data and the geographical (symmetrically, calendar) data (e.g. a radio cell is allocated to an airport or is deployed to serve an important football match) can be 'logically' collected in the 'MNO network topology data' and/or in the 'Geographical data' ('Calendar data') objects. During the project, in addition to the identification and definition of the relevant methods in the pipeline, it will be clarified in which data objects it is most convenient, effective and simple to 'physically' report these pieces of information, for example by evaluating which methods are mostly fed by them. At this stage, we prefer to avoid a definitive assignation of this piece of information to a single data object.

It is important to note that population data will only be used at an aggregated level in the demonstrator scenario of the pipeline, and exclusively in the final step of the pipeline for validation and benchmarking purposes, as well as for grossing up and extrapolation purposes.

We highlight that the methodological framework is modular and is ready to include more sophisticated data integration and fusion methods. However, the elaboration of such methodologies is not the focus of this project, as this will be pursued in a parallel research project funded by Eurostat (ESSnet TSS-METH-TOO, also known as MNO-MINDS).

\ **CONFIGURATION DATA** are used to specify the use case, i.e. selected indicators, time resolution, zoning system, use case specific requirements, etc.

It is important to underline that all the data objects are actually a collection of potentially different data sub-objects with different formats and structures (see **FIGURE 4**), and that they are accessible (i.e. readable) at any point of the high-level pipeline, providing the required information for processing and allowing the subsequent elaborations. An exception is represented by the MNO data, which are ingested at the beginning of the pipeline and elaborated in different intermediate and output data objects.

4.4 HIGH-LEVEL PIPELINE DEFINITION

FIGURE 2 provides a high-level view of the proposed pipeline. The different elements of the pipeline are discussed in the subsequent subsections. Section 4.4.1 Processing network topology data for syntactical checks and spatial cell information illustrates the syntactical checks for network topology data and the module dedicated to the estimation of the spatial cell information. Section 4.4.2 Processing event data introduces the processing of event data, in particular their syntactic checks. These steps are not use-case specific. Section 4.4.3 Multi-scale longitudinal analysis per device describes the multiscale longitudinal analysis at the device level, introducing a few specific elements to concrete use cases (UCs) or groups of UCs. The bottom part of the pipeline, 'Calculation of Indicators', is use case (or group of use cases) specific and collects several modules, from the module 'Aggregation' to 'Output Indicators Quality Checks', and is illustrated in Section 4.4.4 Calculation of indicators per use case and in **FIGURE 5.** This chapter ends with two sections devoted to two crucial topics for this project: Section 4.4.5 Multi-MNO data fusion and MNO-to-NSI data transfer is focused on multi-MNO data fusion and Section 4.4.6 SDC for dissemination and Quality checks of output indicators deals with the privacy aspects of the output of this project.
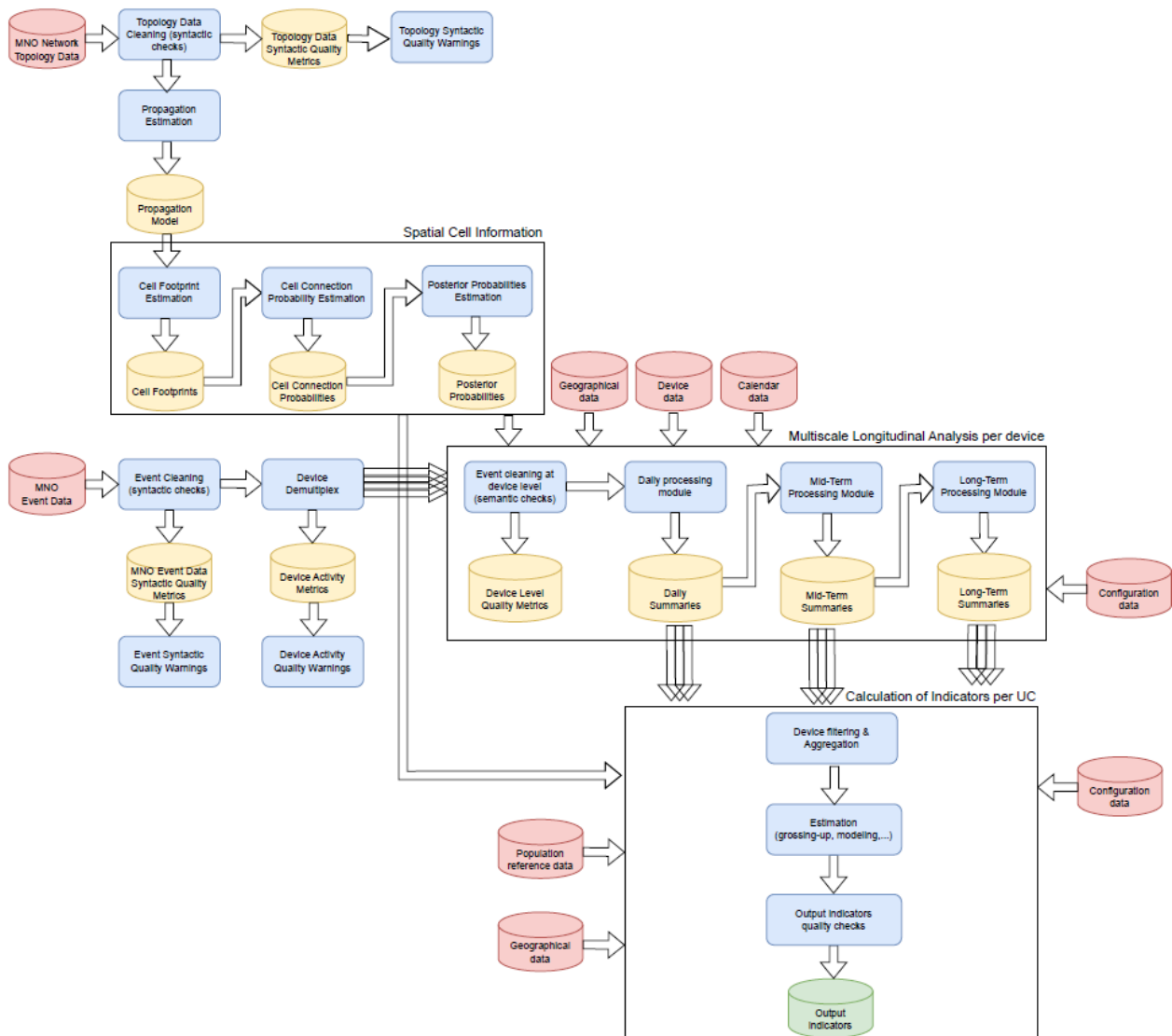
Figure 2: High-level view of the pipeline

4.4.1 PROCESSING NETWORK TOPOLOGY DATA FOR SYNTACTICAL CHECKS AND SPATIAL CELL INFORMATION

In the MNO event $<k,i,t>$, the variable i represents the available information on the location of the device k at time t. It could be the latitude and longitude where device k is found at time t or the cell_id to which the device k is connected at time t. In the latter case, we need to use the topology network data to derive and estimate the geographical areas covered by the cell_id i, and hence the (probable) location of the device. The topology of the MNO network changes slowly and is dynamic. It may include a mobile antenna that changes location daily. Therefore, the topology network data are ideally provided daily by the MNO, so that they always reflect the status of the network. These data are processed daily by the 'Spatial Cell Information' module, which is actually a collection of several sub-modules, as described in detail in Volume III Methods and data objects. The resulting spatial cell information corresponding to each cell_id is crucial to geolocating the MNO event data later in the pipeline.

First, a topology network data cleaning function purges malformed or missing data. A syntactic check is applied, and only valid data can be forwarded to the subsequent phase. At this stage, syntactic quality metrics are produced by reading the data from the topology network data cleaning procedure, in particular the information about the issues found in the data. The quality metrics should not be considered as a standalone object and indeed, in order to maximise their utility, they are in turn read by another function whose task is to launch warnings for specific issues (i.e. quality warnings). Quality metrics and corresponding quality warnings/actions will be described in Task 3, in a dedicated deliverable. Quality metrics and quality warnings are not displayed in the diagram, but are detailed in Volume III – Methods and data objects.

The 'Spatial Cell Information' module is designed to take as input either the coverage area with signal strength values of the network cells directly provided by the MNO or, alternatively, data that allows the estimation of such coverage areas, such as the location of the antennas, their orientation, characteristics (e.g. the technology: 3G, 4G, 5G, etc.), emission power, etc.

In both cases, the output of the 'Spatial Cell Information' module will be a spatial cell information for each cell_id at the given time. The 'Spatial Cell Information' module is described in detail in Volume III, in Section 13.1.

4.4.2 PROCESSING EVENT DATA

The pipeline receives daily[7] flows of event data, each processed according to the following chain:

1. An **event cleaning function** purges malformed or missing data. A syntactic check is applied, and only valid messages can be forwarded to the subsequent phases. At this stage, syntactic quality metrics are produced by reading the data from the event cleaning procedure, in particular the information about the issues found in the data. The quality metrics are in turn read by another function whose task is to launch warnings for specific issues (i.e. quality warnings).
2. The **clean data** resulting from the application of syntactic checks - i.e. through the event cleaning function - is **grouped per device and per day**, generating a separate sub-flow of temporally ordered events for each device. This is done by the 'Device Demultiplex' module. To maintain consistency and keep track of the events after the demultiplex function, a summary table is produced, the 'Device Activity' table. This table includes a device dimension, defined by pseudo-anonymised aliases, and various metrics that will later be used to assess the quality and validity of the data for different use cases. A basic example of such metrics is the number of events per timeslot of a given duration (e.g. one hour), the maximum time between events, etc. This summary table represents a new data object that will reveal its utility for quickly identifying devices with very few events and then filtering them out for some specific use cases. This will allow extreme anomalies in the data (e.g. devices with an extremely high number of events in the fixed timeslot) to be captured for subsequent consistency checks and to speed up the computation in later phases.
3. Each sub-flow is then subject to a **multi-scale longitudinal analysis**, detailed in Section 4.4.3 Multi-scale longitudinal analysis per device. The spatial and temporal scaling can be determined by the use case described in the Configuration Data use case.

---

[7] Subject to agreements with MNOs. The ingestion of daily flows of event data is the ideal scenario.

4.4.3 MULTI-SCALE LONGITUDINAL ANALYSIS PER DEVICE

This section describes the parallel analysis of individual device events, at the core of which is the processing of daily batches of event data in a multi-scale dimension, briefly depicted in **FIGURE 3**. For better clarity, the different functions and data objects included in the process are described in the following paragraphs.
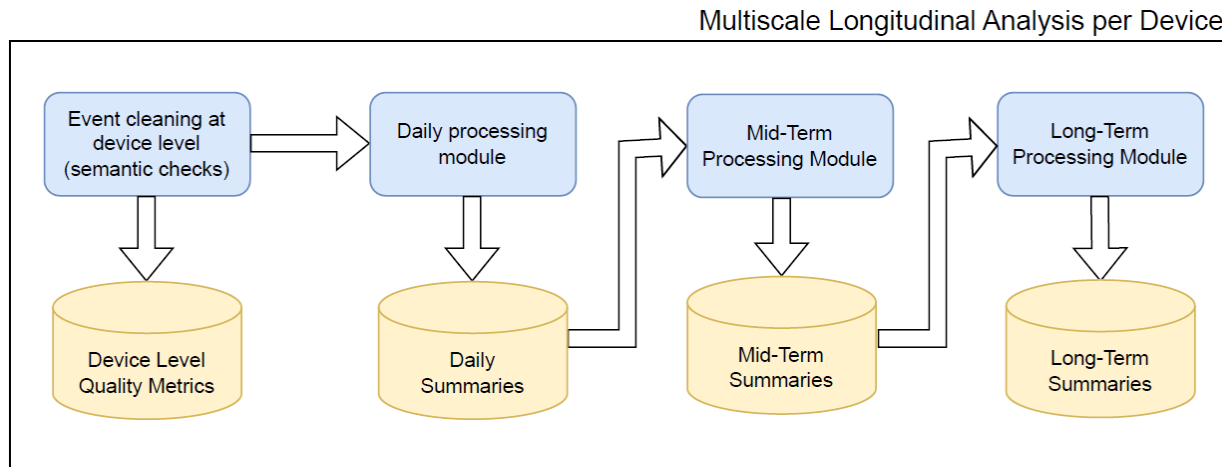

Figure 3: Multi-scale longitudinal analysis at the device level (in the pipeline processing)

\ EVENT CLEANING AT DEVICE LEVEL – *SEMANTIC CHECKS (FUNCTION)*

The individual data flows pass through an event cleaning function, which differs from the syntactic cleaning function described earlier. In the first cleaning event function, before the demultiplex function, malformed or syntactically incorrect records were purged. In this step, the individual flows are cleaned according to semantic logic. Data and events may be apparently correct, but still contain defective information that deserves some attention. This is the case, for example, of non-existent cells, i.e. cells whose information appears to be correct, but which in practice do not correspond to any cell_id entry in the network topology object or cells whose id follows the correct format, but with an invalid or meaningless value, e.g. the null value. Another important semantic error to detect at this stage is the case of two adjacent (in time) events with corresponding cell_id formally corrected, but separated by a geographical distance that is incompatible with any means of transportation in the reported time interval. This type of error is identified and dealt with in this step.

\ QUALITY METRICS *(OBJECT)*

After dealing with the individual event flow issues, the event cleaning step module produces a new 'Quality Metrics' object. These metrics differ from those elaborated in the first part of the flow since they are computed at an individual level for each device. Of course, the metrics can be customised according to the use case or the criticalities that may need to be investigated. Furthermore, given the nature of the data and the networks, a certain quota of problematic aspects has to be expected. Therefore, for some quality measures, it could be useful to establish a specific error threshold. Quality metrics and corresponding quality warnings/actions are described in Volume III – Methods and data objects.

\ DAILY PROCESSING MODULE *(FUNCTION)*

The main processing module of the individual flows is the 'Daily Processing' module, which is carried out immediately after the event cleaning at device level. The data is read and processed in daily batches for which some temporal aspects have to be defined. In particular, it will be useful to have some continuity between adjacent daily processes (it could be some kind of overlap or another different way) so that the last messages of the last daily batch are still connected to the first messages of the new daily batch, and the information occurring at the

temporal boundary of the batch is not lost. In addition, it could be useful to fix the start time of the batches around the time of least load of the network (3-4 am), rather than at midnight. All these aspects will be carefully defined in the agreements with the MNOs.

The 'Daily Processing' module aims to compute, on a daily basis, all the syntheses of the individual data that are required as intermediate results to obtain the statistics of interest in the various use cases. An example of this synthesis produced by the 'Daily Processing' module is the Default time segmentation (for further details, please refer to the description of the method in Volume III – Methods and data objects, Section 13.3 Continuous time segmentation). In any case, the 'Daily Processing' module will support the availability of more than one single method, also in terms of implementation. A candidate method to be implemented in Task 4 (software development) together with the Continuous Time Segmentation is the algorithm proposed in Volume III in Section 13.2 - Daily Permanence Score Estimation. Its usage is proposed and methodologically justified in Volume II in Chapter 4 – M-Usual Environment Indicators Use Case. These two methods, which reflect the requirements for the longitudinal analyses-type of UCs[8], will be both implemented in the pipeline in Task 4 (software development) as independent modules or sub-modules, with no interdependencies between them. The parallel/alternative implementation of these methods is required since it is directly linked with the UCs these methods serve and as well as an example of the availability of multiple methods for one single module.

Important remark 1

Since the pipeline should ensure flexibility, evolvability and generality, the methodological design of the 'Daily Processing' module will allow for a (limited) set of potentially different functions (sub-modules) that process the daily data flow and produce a set of different daily summaries. In principle, each daily summary can serve a different use case or group of use cases, and be stored in a different data sub-object. At the moment, it is not yet clear whether a single (default) time segmentation sub-module will suffice for this project to serve all use cases (preferred option), or whether a second or third segmentation logic will need to be implemented (non-preferred but still allowed option). This will become clearer as more details on the use case definition and related methods become available as the project progresses. However, even if during this project a single Daily Processing sub-module will suffice for all use cases considered for demonstration in the project, the reference pipeline will be designed and implemented to also serve future projects and needs of other NSIs for which we cannot exclude a priori the possibility to compute alternative daily summaries. All the data objects produced by the daily processing module are collected in daily summaries, which are described in the following paragraph.

Important remark 2

It is worth noting that in the multi-scale analysis, the data flow has an univocal direction from smaller to larger scales, in line with the principles of 'multiscale longitudinal analysis', the 'input data accessibility from any point of the pipeline' and the 'bottom-up one-way processing' stated at the beginning of the chapter. In other words, analysis at a smaller-scale provides input to the analysis at a larger scale (e.g. from daily to mid-term, and from mid-term to long-term, or possibly even directly from daily to long-term). Once written, the summary at one timescale may be read and used by any module, including processing functions at a smaller timescale. However, the pipeline design does not allow the summary at one timescale to be influenced (written or updated) based on information from a larger-scale (e.g. the function that writes daily summaries cannot take input from mid-term or long-term summaries).

---

[8] In addition to the Continuous Time Segmentation and the Daily Permanence Score which are methods that reflect the requirements of longitudinal analyses-type of UCs, the Present Population Estimation method - which reflects the requirements for the 'snapshot' type of UCs - will also be implemented in the software pipeline in parallel.

This modular structure of functions and objects produced in the individual longitudinal analysis ensures the extensibility of the pipeline in case of new emerging needs related to new case studies. The modularity of the (set of) functions and (set of) objects is represented in **FIGURE 4**.
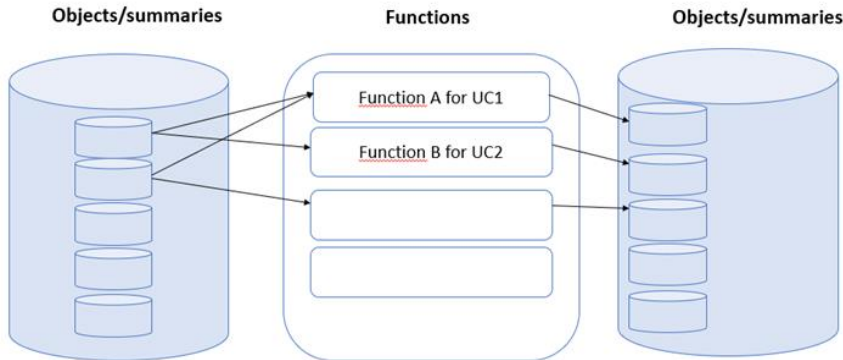

Figure 4: Representation of the modularity of functions and objects. Data objects (cylinders) and data functions (rectangles) are composed by multiple sub-objects (smaller cylinders) and sub-functions (smaller rectangles). Each data sub-object can be an input for multiple sub-functions.

The diagram shows:

- an exemplification of data sub-objects/summaries belonging to the daily summaries (the data object), represented by the smaller cylinders within the big cylinder on the left;
- the usage of these summaries by different sub-functions of the mid-term processing module, to serve potentially different purposes for different use cases, represented by the arrows and indicating that the data sub-objects are used as input by the sub-function represented by the rectangles in the middle; and finally,
- the sub-functions that produce new data sub-objects, i.e. the mid-term summaries, represented by the smaller cylinders on the right, are ready to be used in the subsequent steps of the pipeline.

Important remark

*It is important to clearly distinguish the semantic level of the pipeline from the implementation level. The pipeline description is **not** an implementation proposal, i.e. the 'Daily Processing' module should not imply that the software must run daily. The 'daily reasoning' at the semantic level does not mean that the data must be ingested daily at the implementation level. They can be ingested at shorter or longer intervals – this is purely an implementation detail, and the periodicity of ingestion (at the implementation level) should be seen as distinct from the periodicity of logical processing (at the semantic level). More specifically, the term 'daily' designates the reference period of the intermediate results produced by the 'Daily Processing' module, regardless of whether the data are provided in 10-minute or weekly batches, and regardless of whether the software runs every day or once per week in order to provide these daily summaries. The periodicity of data ingestion, as well as the periodicity of software runs, are 'configuration' settings that go beyond the pipeline description at this stage. These will be discussed and determined during the implementation tasks.*

\ **DAILY SUMMARIES** *(OBJECTS)*

The 'Daily Processing' module will provide a set of daily summaries, i.e. a synthesis of the information gathered from the events associated with each mobile device as necessary for the target use-cases. It should be noted that, in line with the principle of data minimisation in the GDPR, information that is not strictly instrumental to the implementation of the target UC will not be included in the daily summaries (for example, if the target UCs only require the measurement of the cumulated amount of time spent in one place during the day, the more detailed information about the exact start and end times of each stay will not be stored).

As already mentioned in the previous paragraph, an example of this synthesis produced by the 'Daily Processing' module is the Default time segmentation (for further details, please refer to Section xxx Continuous time segmentation in Volume III – Methods and data objects). Another example of a daily summary is the permanence score, i.e. the estimate of the time spent on a daily basis for the UC on the M-Usual Environment Indicators (for further details, please refer to the pipeline application to this UC in Volume III, Annex III and to the definition of the UC in Volume II, Chapter 4). The daily summaries might not be intended to be limited to these examples. The exemplification of the pipeline's application to other use cases, which will be carried out in the upcoming project phase, will allow us to assess whether there is a need to add different summaries. In any case, the pipeline design will be implemented in such a way that it is easily evolvable and can be enriched with additional sub-modules implementing different logic, should additional needs arise.

\\ MID-TERM PROCESSING AND LONG-TERM PROCESSING MODULES *(FUNCTIONS)*

One of the main potentials of MNO data is related to the **longitudinal information** they provide for each device. In order to properly exploit the longitudinal information, we propose a multi-scale longitudinal analysis in which the individual device daily summaries are further analysed, first at the mid-term temporal scale and then at the long-term scale.

- The purpose of the 'Mid-Term Processing' module is to reduce and synthesise the daily summaries into mid-term summaries. Individual device daily summaries are analysed on a mid-term scale (e.g. a month, a period of several weeks, a quarter, etc.). As in the case of the daily summaries, the mid-term summaries are intended to be a collection of several mid-term profiles connected to different specific use cases and reusable by future use cases.
- The 'Long-Term Processing' module is designed in a similar way with the aim of capturing the long-term behaviour of the device. Daily and mid-term summaries of individual devices are analysed on a long-term scale (e.g. over a period of several months).

The 'Mid-Term Processing' module takes individual daily summaries as input, analyses them further and produces mid-term summaries. The mid-term dimension can be a solution to consider seasonality and multi-day activities (as multi-day trips). The mid-term summaries are conceived as a collection of several mid-term aggregated outputs related to the specific use cases and reusable by other use cases if they also fit the other purposes. The organisation of the mid-term function and data object as a modular set of several functions and data objects facilitates the evolvability of the pipeline as new needs arise.

Important remark

There are cases where the need for mid-term outcomes (requiring a mid-term processing) is quite clear; e.g. in the case of tourism statistics, where many outputs are produced on a monthly basis. In other cases, e.g. for the identification of usual residence/home location, it seems that the mid-term data processing can be skipped and one can pass directly from daily processing to long-term processing. To ensure generality, we prefer to keep the mid-term processing, without excluding the possibility of eliminating it if the pipeline application to all the use cases considered in this project does not highlight the need for it. It is also worth mentioning that the mid-term processing brings other benefits, for instance, for compliance with GDPR: it allows the consolidation of the individual information encoded in daily summaries and reduces the information that is stored for a longer period. This is in line with the principle of storage limitation and data minimisation of the GDPR. The choice of reference timescales (e.g. daily, weekly, monthly, quarterly, annually) is among the design aspects that need to be agreed upon at a European level to ensure consistency and comparability. This project will propose these aspects for discussion to the ESS Task Force on the use of MNO data for Official Statistics (TFMNO) in order to facilitate the standardisation at European level. As a first consideration for the discussion within the TFMNO, we can set the value for the mid-term processing to one month, in order to accommodate many statistical outputs produced on a monthly basis. If this default value is supported by the TFMNO, and adopted at the European level, it would be useful to decouple the

functional purposes of the mid-term processing from its strategic use in terms of privacy guarantees. For instance, it could be decided to standardise the data processing and storage at a European level by aggregating daily summaries into monthly summaries, and by allowing them to be stored for one quarter, after which the daily summaries could be deleted. The same reasoning applies to the mid-term summaries; they have to be stored until the mid-term summaries are calculated and can then be deleted.

The 'Long-Term Processing' module is designed in a similar way, with the aim of capturing the long-term behaviour of the device. For instance, on a long-term scale, we will define functions that will analyse previously computed summaries to identify the home location, the usual environment and any other labelled/tagged information required by those UCs that require a long-term observation of the device.

The multi-scale analysis can be based on all the summaries produced by the previous steps, and the information stored in the data objects described at the beginning.

\ MID-TERM AND LONG-TERM SUMMARIES AND PROFILES *(OBJECTS)*

The output of the 'Mid-Term Processing' module will include, for example, the most frequent place of overnight stay of a user in a certain month or season of the year, which may not necessarily correspond to their main place of residence. As mentioned above, some UCs will require the production of outputs related to the mid-term interval, such as the tourism statistics UC.

The output of the 'Long-Term Processing' module will include, for example, anchor points (e.g. home location, usual environment) and other labelled/tagged information that is required by the use cases and requires a long-term observation of the device.

Box 1: Concise overview of the multi-scale summaries

Using the intermediate results provided by the 'Event Cleaning' module and integrating them with the available information on cells, places, calendar and users, three main levels of summary objects can be identified:
- **Daily summaries:** provide information on the events that occurred at different time slots during each single day for each device.
- **Mid-term summaries:** aggregate the daily information at specific mid-term intervals (e.g. monthly) to produce summaries relevant for specific use cases.
- **Long-term summaries:** provides the corresponding summaries for long-term intervals (e.g. annually) relevant to specific use cases, such as M-Home Location Indicators or M-Usual Environment Indicators.

Each summary object is built by the module that corresponds to the appropriate time interval. Additional modules could deal with the processing of the information at different temporal resolutions and provide information for related summary objects. It is important to note that in this high-level description of the pipeline, we do not assign a specific format to the summary objects, while this aspect concerns the semantic and syntactic levels of the pipeline where the format of the objects and sub-objects will be specified to ensure the standardisation. They are therefore referred to generically as 'data structures'. The exact format will be defined later in the project.

Each summary object may need to be integrated with different types of contextual information. For example, the calendar information, will most likely be needed in the mid-term and long-term summaries, while the geographical reference data will already be used in the daily summary.

The organisation of the mid-term and long-term functions and data objects as a modular set of several functions and data objects will facilitate the evolvability of the pipeline as new needs arise.

4.4.4 CALCULATION OF INDICATORS PER USE CASE

The daily, mid-term and long-term summaries at the individual device level, processed by each single MNO are now ready to be elaborated and aggregated across the population of mobile devices (or subgroups thereof), first within each single MNO and then across different MNOs, to produce the final statistical indicators. This set of modules aims to specify the different steps (i.e. sub-modules and relative methods) required to produce the final specific use case indicators. These clarify which steps are basic components of our pipeline (i.e. they must be considered in the reference scenario), and which ones will only be included in the implementation of this project for the sake of demonstrating the pipeline itself (i.e. they are functional to the demonstrator scenario).

The basic logic steps to produce the final estimates in the reference scenario are summarised in Box 2:

Box 2: Ideal steps of the estimation phase

1. Filtering/selection and aggregation of device data within the single MNO at the finest standardised geographical level; i.e. the regular 100x100 m grid, as in the M-Usual Environment Indicators use case and other use cases requiring longitudinal observations, or the single cell (or cluster thereof) as in the present population use case.

2. Merging of the single MNO aggregates into multiple MNO aggregates at the finest standardised geographical level, i.e. the regular 100x100 m grid.

3. Projecting the multi-MNO aggregates from the finest level to the geographic unit systems relevant to the use case.

4. Estimating the target population indicators (this might require: grossing up procedures based on weights or statistical models capable of producing target population estimates without weighting, using the MNO data as the primary source, or just simple unweighted data, if the multi-MNO results are used as auxiliary information in statistical models).

5. Applying SDC methods to the final output before dissemination.

These steps are not always fully separable and many adaptations to the specificity of the UCs can be envisaged. For instance, in the Present Population use case described in Chapter 3 in Volume II – Use cases, steps 3 and 4 can be performed together and the final estimates will consist of the spatial density of the present population broken down into the required zoning system. Similarly, in the M-Usual Environment Indicators use case described in Chapter 4 of Volume II – Use cases, the first step does not require any filtering of the relevant device data, since this is already included in the labelling process of the long-term analysis and only the aggregation of the device data within the single MNO is required before proceeding with the following steps.

In addition, it is remarkable that in this pipeline we propose the spatial aggregation (step 3) before the final estimation (step 4), mainly because the external auxiliary information needed in the estimation phase is often not available at the finest standardised geographical level (the regular 100x100 m grid) at which we can compute multi-MNO aggregates, so we must first further aggregate the multi-MNO summaries to the largest level and then apply weighting procedures or other methods to estimate our target statistical output. Indeed, this is the current situation if we assume that sample survey and administrative data will be used to integrate MNO data. If other data sources should be used to integrate MNO data for the production of target statistical indicators, and these additional data should be available and reliable at the finest level at which we can aggregate multi-MNO data (the regular 100x100 m grid) the steps 3 and 4 could be swapped. However, the purpose of landscaping potential non-MNO sources to be integrated with MNO data and proposing advanced methodologies for the integration is beyond the scope of this project. Therefore, we have chosen to design the pipeline meeting first with the configuration data scenarios that seem most likely. The exchange of these two steps can be seen as an extension of the pipeline for future evolution.

It is important to note that the kind of aggregation and the geographical and temporal resolution of the data to be aggregated can differ from one use case to another. This flexibility is easily provided by our pipeline, where the 'calculation of indicators' module is also composed of different functions that can access all the previous device summaries computed at different temporal resolutions (i.e. daily, mid-term and long-term). The representation of the 'calculation of indicators' module is provided in **FIGURE 5**.

It is worth mentioning that not only the methods and external data used in the 'calculation of indicators' module, but also the set of MNO device data might depend on the use case at hand. In other words, the observed population to be considered in the aggregation might differ across use cases. For instance, we might decide not to use the devices for which we have 'too little data' or 'too sparse observations', where the threshold that defines 'too little' and 'too sparse' might vary from one use case to another. For these reasons, a 'filtering/selection' module is considered before the aggregation, in order to focus only on the units (devices) that are relevant to the use case at hand.

It should be noted that potential differences in the absolute figures might result from the different sets of observed statistical populations targeted by the different use cases. Following the practice in official statistics, such discrepancies are resolved in the estimation step (via some inferential method, or by rescaling and reweighting, or by providing only relative figures).
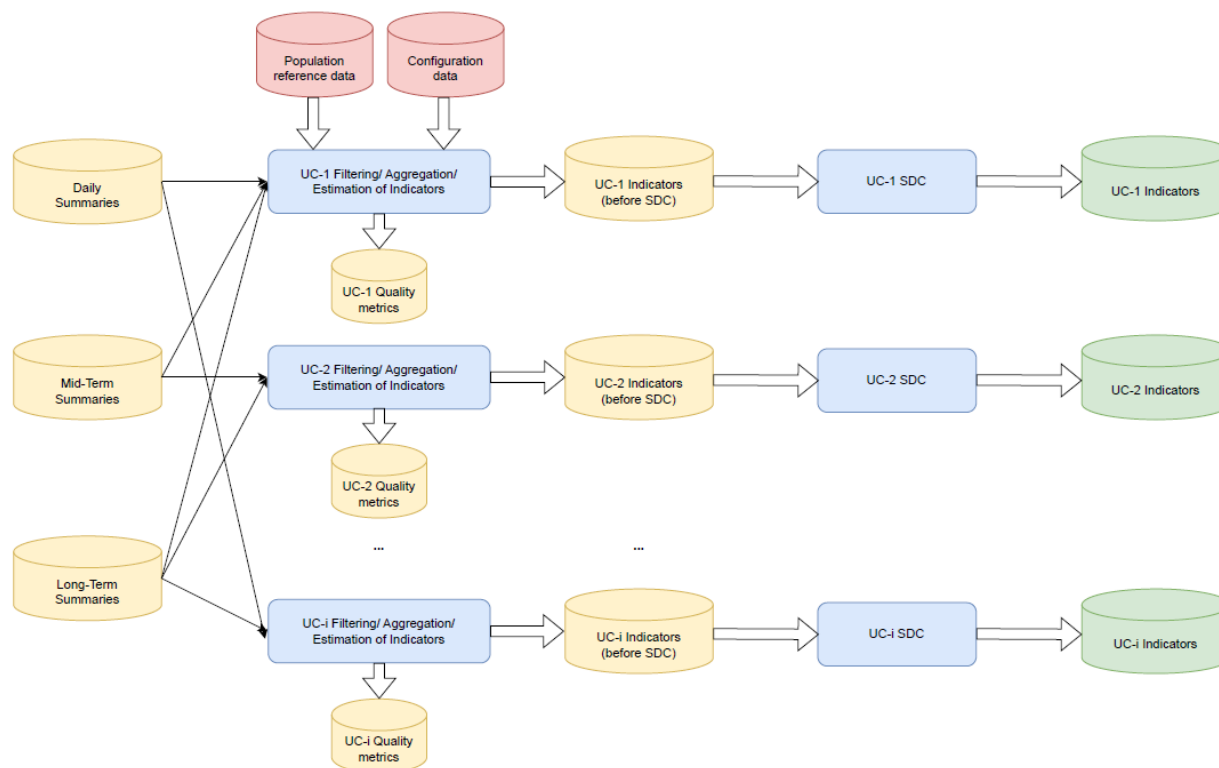

Figure 5: Representation of the integration and aggregation process

4.4.5 MULTI-MNO DATA FUSION AND MNO-TO-NSI DATA TRANSFER

It is important to note that these are the pipeline modules which, in the hypothetical reference scenario, can potentially be performed partly outside the MNO premises. In fact, the previous steps for the calculation of the statistical indicators can be performed by different actors in different premises. In our demonstrator scenario, due to the current privacy legislation, the first step of the process, 'aggregation of device data within the single MNO at the finest geographic level' will take place on the premises of the MNO, and only aggregated data will be made

available outside the single MNO. In addition, since aggregated data can also be at risk of reidentification, in the demonstration scenario and in the testing phase of the pipeline we will apply a SDC method in order to ensure that the risk of reidentification of the single MNO aggregates is properly limited. The application of this SDC method on the single MNO aggregates is strictly related to the pipeline's testing purpose on real MNO data. It is envisaged for compliance with the current privacy legislation and under the specific agreements with the partner MNOs of this project in the testing and demonstration phase.

From the second step, the 'combination of the single MNO aggregates into multiple MNO aggregates', the NSI itself or a trusted third party or jointly recognised secure enclave will be the premise for the data elaboration in the reference scenario.

In addition, it is important to note that in these pipeline's modules, the MNO data can potentially be integrated with other non-public data sources that are available to the NSI.

Due to the importance and sensitivity of these aspects, during the project we will apply the previous steps in the way they comply with the current privacy legislation and the specific agreements with the MNOs. In the case changes to the sequence of the ideal steps are needed, they will be strictly related to the purpose of testing the pipeline with real MNO data. Therefore, they will be marked and attributed to the demonstrator scenario. In fact, these steps represent the part of the pipeline where the reference and demonstrator scenarios might differ most, as they scope what MNO data can leave the MNO premises and be integrated with other data sources. Furthermore, the release of the MNO data to the NSI, as well as their integration/combination with other data, are the functions most affected by the business and privacy sensitivity. Therefore, the conditions under which the project operates in the demonstration phase may have a direct impact on what can be implemented and tested.

By **basic approach for Multi-MNO data fusion** we refer to the combination of the previous factors, namely:

- Only single MNO aggregated data are available for multi MNO data fusion;
- Single MNO aggregates are additionally processed with SDC techniques before leaving the single MNO premises.

The basic approach is fully compliant with current privacy-related regulations and the specific agreements with the partner MNOs of this project. Hence, it will be implemented and tested in the demonstrator scenario. The orange rectangle in **FIGURE 6** highlights delimitate the modules of the pipeline which apply the Multi-MNO data fusion and the MNO-to-NSI data transfer in the basic approach.
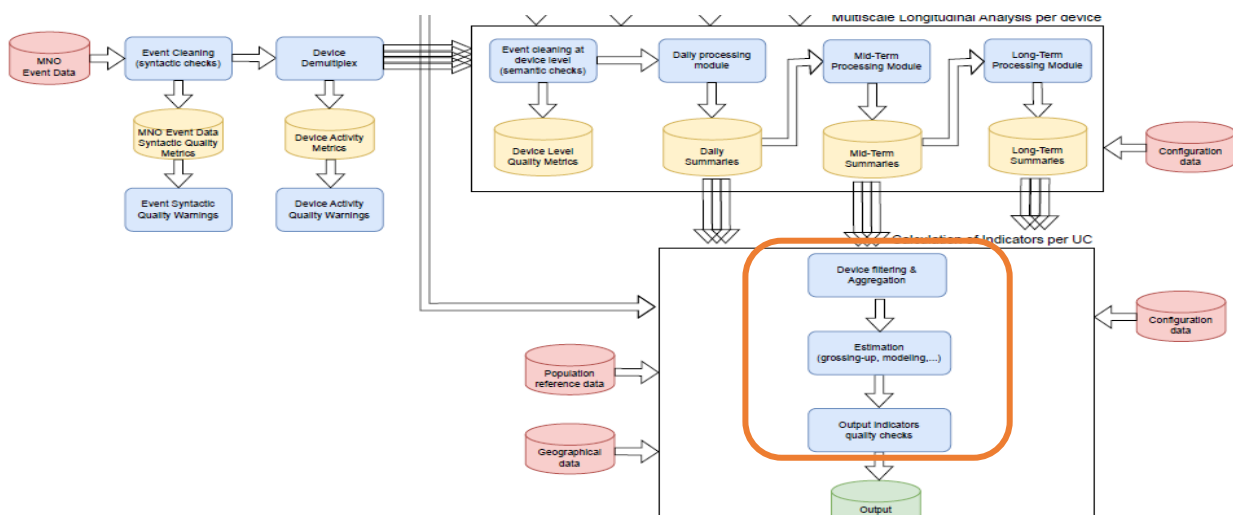

*Figure 6: Representation Multi-MNO data fusion and the MNO-to-NSI data transfer in the **basic approach for Multi-MNO data fusion***

It is worth noting that the second condition of the basic approach can be relaxed, i.e. the application of SDC methods to the single MNO aggregates, can be relaxed and SDC methods avoided for single MNO aggregates if the combination of single MNO aggregates into multi-MNO aggregates takes place in a secure enclave equipped with privacy enhancing techniques.

However, we aim at designing a pipeline that is not limited to the demonstrator scenario, but which can serve more general situations, driven by the methodological needs of standardisation for the whole ESS, under potentially more favourable private-public partnership conditions, and evolvable towards the most advanced methodological solutions. For this reason, we refer to an **advanced approach for Multi-MNO data fusion**, in which potentially single MNO data can be fused into Multi-MNO data at the individual level in a secure environment equipped with privacy enhancing techniques and, subsequently, jointly processed. It is crucial to specify that in the **advanced approach** we envisage, it is only expected the fusion of individual data (for instance, daily summaries) and not of any nano-data, i.e. the stream of the MNO events generated by the single mobile network. The fusion of any nano-data is not expected neither in the case the advance approach would operate in a secure environment. This is in line with the principles of data and storage minimisation that animate the whole data processing pipeline.

Since tools for privacy enhanced techniques are not included in this project, and given the current agreement with the partner MNOs, the advanced approach will not be implemented, nor demonstrated. Nevertheless, it is relevant noting that the designed pipeline is fully compatible with the advanced approach for multi-MNO data fusion. The modules and sub-modules already included in the pipeline fit both the basic and advanced approaches, the main difference is that in the advanced approach some modules are required to run in a secure environment. An example of the modules that are requested to run in a secure environment in the advanced approach is provided in **FIGURE 7** highlighted with a light blue rectangle. As anticipated, only aggregated individual data are required to contribute to the data fusion. In the example the daily summaries from the single MNO contribute to the mid-term summaries for multi-MNO, while the single MNO event data are still processed by the single MNO in their own environment.
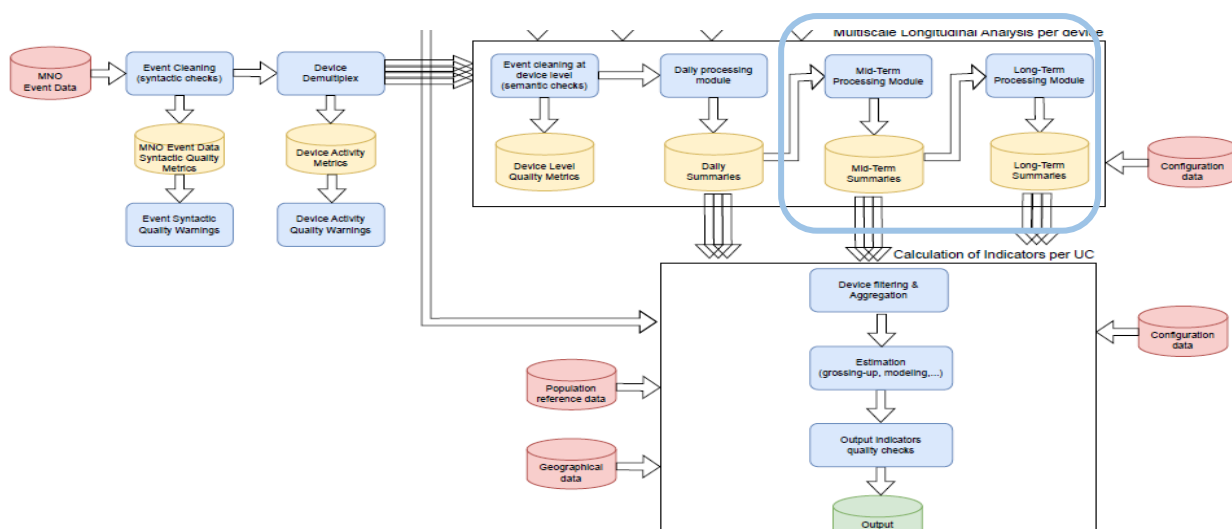

*Figure 7: Multi-MNO data fusion in the **advanced approach***

Even if the advanced approach for data fusion cannot be implemented in this project, we envisage it as an evolution of the current situation, keeping in mind that it will enable not only multi-MNO data fusion, but also integration at individual level with other data sources that can empower the production of the statistical output indicators, greatly improving the estimation phase, i.e. the transformation of the device-based statistics into the target statistical population or population of interest. Finally, an advanced approach to data fusion can be especially relevant for some use cases, e.g. the inbound tourism statistics, where the combination of separate aggregates from several visited MNOs can produce biased results in case of cross-roaming, as illustrated in Chapter 12 – Inbound Tourism in Volume II - Use cases.

Finally, it is worth recalling that there is a separate parallel project investigating methodological developments for the integration of MNO and non-MNO data. This project can propose the definition of new approaches and scenarios that are not available during our implementation phase. For this reason, for all the steps mentioned in **BOX 2**, we select some basic methods that are described in Volume III – Methods and Data Objects and further implemented.

In the end, it is still useful to point out that this project is not intended to formalise a proposal on which components of this pipeline should run at the MNO premises and which outside the MNO premises.

4.4.6 SDC FOR DISSEMINATION AND QUALITY CHECKS OF OUTPUT INDICATORS

The quality of the final output is assessed, and quality indicators are produced in line with the practice in official statistics. In addition to the quality checks, in the reference scenario, SDC methods will be applied before the release of the final statistics aggregating data over multiple MNOs, as is usual in NSIs procedures. However, SDC methodologies for MNO data might be a complex issue, and specific SDC techniques can be applied for public release of the results, while different ones can be required for the release of scientific-use files. Both are beyond the scope of this project. In addition, no data will be publicly released from the testing and demonstrating phases of this project based on real MNO data, being the dissemination stage out of the scope of this project.

Hence, in the reference scenario, we can envisage the exploitation of privacy enhancing in-processing measures for the fusion of multi-MNO data and the application of SDC methodologies before the dissemination of the results. The processing and output privacy requirements in the reference scenario are depicted in **FIGURE 8**. However, none of the previous ones are in the scope of this project and will not be designed, nor implemented, in the demonstration scenario.
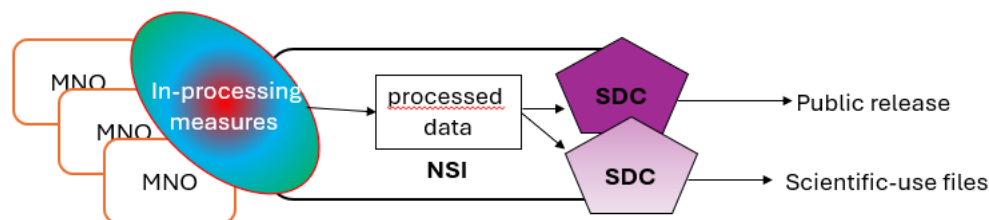

Figure 8: Processing and output privacy requirements in the reference scenario

On the opposite, in the demonstrator scenario exclusively, as already introduced in the previous section, the intermediate aggregated data from each single MNO will be additionally processed using SDC methods, and before being combined in multi-MNO aggregates and accessed by the NSI. This, in order to comply with the current privacy legislation, which in some countries may require the 'anonymisation' of the data leaving the MNO premises, as explained in Chapter 3 Reference scenario. This requirement is illustrated in **FIGURE 9** and its solution/application in the demonstrator scenario is further described in Volume III – Methods and data objects.
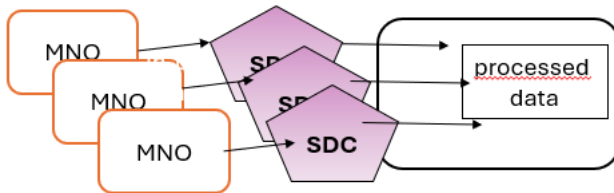

Figure 9: Processing privacy requirements in the demonstrator scenario

5 CONCLUDING REMARKS

This document defines the methodological framework proposed by the Multi-MNO project for the processing of multiple MNO data for official statistics. It comprises the project's conceptual framework and high-level requirements, the definition of the reference and demonstrator scenarios and the main details of the data processing flow.

As a brief of the high-level definition of the pipeline, we distinguish three main workflows for the data processing, namely:

- The advanced geolocation workflow;
- The longitudinal analysis of individual device data workflow;
- The aggregation and estimation methods workflow.

The proposed pipeline architecture is regarded as sole basis for any future development or evolution path for the processing of MNO data for official statistics, in the future reference scenario. The pipeline architecture presents as main novelty and cornerstone in the data processing, the multi-scale temporal workflow (i.e. the longitudinal analysis of individual device data). The methodological design of this workflow also reflects the alignment with the core principles of GDPR on data minimisation and data storage.

The project is only one part of a series of initiatives undertaken by Eurostat for the re-use of MNO data for official statistics[9]. Therefore, it is limited in scope and, at the same time, the proposed reference scenario is dependent upon the development of the other related initiatives and subsequent institutional and public-private agreements. In the current document, we carefully explain in the high-level pipeline definition, the aspects and/or workflows that are outside the scope of our project.

This document is the first of three volumes, which altogether build the content of the project deliverable D2.2 – Updated version of technical documentation for scenarios, requirements, use cases and methods, and high-level architecture. Volume II of this publication defines a set of use cases / statistical indicators which can derived from the processing of MNO data, while Volume III specifies in detail the full set of methods and data objects developed for four use cases (namely, Present Population Estimation, M-Usual Environment Indicators, M-Home Location Indicators and Internal Migration).

---

[9] See position paper: Reusing mobile network operator data for official statistics: the case for a common methodological framework for the European Statistical System – 2023 edition - Eurostat (europa.eu)

ANNEX 1 – OVERVIEW OF PROJECT TASKS

The high-level objective of the Multi-MNO project is to develop a complete, open end-to-end processing pipeline for the production of future official statistics based on MNO data, and to demonstrate it across data from multiple MNOs. The term 'processing pipeline', in this specific context, designates a combination of a fully documented methodological and quality framework, plus the implementation of a reference open-source software pipeline compliant with the said framework.

To achieve this, the following tasks are planned:

- Task 0 – Setup an agile collaboration tool
- Task 1 – Setup and maintenance of a public website
- Task 2 – Definition of scenarios, requirements, use-cases and methods
- Task 3 – Architecture, business processes and quality framework
- Task 4 – Open-source software implementation and documentation
- Task 5 – Deployment and execution of tests on real data from multiple MNOs

From the tasks listed above, Task 2 to Task 5 are purely technical, and shape the concrete content of the 'processing pipeline'. More in detail, the scope of each of these tasks is:

- **Task 2**: define and detail the reference scenarios, the high-level requirements of the methodological framework, as well as a limited set of use-cases (between 6 to 12 distinct statistical end products) and the high-level architecture of the processing pipeline.
- **Task 3**: develop a stand-alone quality framework specific for MNO data, and define the business processes that enable the application of the methodological framework and the open-source software pipeline.
- **Task 4**: implement, in open-source software, a reference processing pipeline adhering to the requirements and specifications developed in Task 2 and Task 3.
- **Task 5**: conduct at least two test runs with data from the five subcontracting MNOs (A1 Slovenia, Orange Spain, Post Luxembourg, Vodafone Italy and Vodafone Spain).

All project deliverables will be publicly disseminated on the project website Multi-MNO project | Eurostat CROS (europa.eu) (in the context of Task 1).

ANNEX 2 - CONCEPTS AND DEFINITIONS

In this annex, we build a common language between industry and official statistics. To this end, we introduce and define the concepts the project is based on.

Whenever possible, the project builds on existing, widely accepted concepts and definitions in the domain of official statistics, but we also introduce a number of new concepts and definitions whenever existing ones are not satisfactory or not yet adequately consolidated to deal with the processing of MNO data. This section is crucial to build a common language between industry and official statistics.

This annex is conceived as a living glossary that can be enriched each time the project encounters the need to define and clarify the meaning of new or ambiguous concepts. Therefore, the glossary is a standalone output of the project itself and not exclusively linked to the current deliverable, being relevant to all other project deliverables.

Table A.1: Concepts and definitions

CONCEPT	DEFINITION
Aggregation	Aggregation refers to the process of combining individual data points into a summarised representation. It involves applying functions like sum, count, mean, etc. to condense and analyse the data. This can be accomplished with a single variable or across multiple variables (or dimensions).
Call Detail Records (CDR) and Internet Protocol Detail Records (IPDR)	Call Detail Records (CDRs) and Internet Protocol Detail Records (IPDRs) are records that capture information on the telecommunication transactions performed by a mobile phone. CDRs are generated every time a subscriber uses services such as calling (in/out) and text messaging (in/out). IPDRs are generated every time a subscriber is accessing the Internet. CDRs and IPDRs include the following information: • type of telecommunication transaction (voice, text, data); • who made the telecommunication transaction (phone number/IMSI of the caller); • who received the telecommunication transaction, if applicable (phone number/IMSI of the recipient); • the date and time the telecommunication transaction was made; • the duration of the telecommunication transaction, if applicable; • location of the caller (e.g., mobile network cell) at the time of the call; • location of the recipient at the time of the call, if applicable; • and typically, dozens of usage and diagnostic information elements (e.g., reason for telecommunication transaction termination, any fault condition encountered, etc.). CDRs and IPDRs are collected on a regular basis for the generation of telephone bills. Additionally, they can be used for other purposes, such as analysing network performance and producing usage, capacity, performance and diagnostic reports. Source: ESSnet Big Data II project

CONCEPT	DEFINITION
Cell_ID	Identifier of a mobile network cell following [CGI and eCGI standards](#) Note that some operators might use a different term internally.
Data objects	Organised collection of data used to store information or used by functions to produce information. A data object can be a collection of different sub-objects.
Statistical data producers	Authorities responsible for the collection and compilation of statistics. It mainly includes the National Statistical Institutes (NSIs) and Other National Authorities (ONAs) and Eurostat (i.e., ESS partners). Their main goal is to obtain a set of standardised reference methods and tools adhering to the requirements and principles of statistical production, which allow them to realise the potential of MNO data for the production of official statistics. Additionally, they are also interested in establishing collaboration mechanisms with MNOs that pave the way for sustainable access to data in reasonable conditions. *In the context of this project, the concept should not be used interchangeably with 'data producer', which is a broader concept, that designates any entity that generates and creates data (not necessarily official statistical data) as an output.*
Data providers	From a traditional perspective, this category of stakeholders of statistics includes the respondents to surveys, but also national institutions that are keepers of administrative data files such as the Social Security or Tax Administration authorities. In the context of this project, the data providers' category comprises the MNOs. In recent years, many MNOs have launched initiatives aimed at developing new value added products and services to monetise their data. MNOs are interested in contributing to society by sharing their data, but they also want to protect their legitimate commercial interests, so in order to incentivise their participation in this and other future initiatives, the proposed approach shall look for a reasonable trade-off between both aspects. An incentive for their participation is to gain privileged access to the knowledge and the data processing software developed by the project.
Statistical data users	Within this group of statistics stakeholders, it can be distinguished between institutional users (policy makers) and other external users such as researchers, private companies, citizens, and mass media. As future consumers of the statistical products to be developed in the project, their main interest is that these products provide detailed, reliable, up-to-date information relevant to their goals and responsibilities (e.g. policy makers will be interested in new statistical products with the potential to improve policy assessment and decision making).
European Statistical System (ESS)	The European Statistical System (ESS) is a network of NSIs and ONAs responsible for producing official statistics in EU Member States (MS). The ESS ensures high-quality, comparable statistics for evidence-based decision-making in the EU. Eurostat coordinates and supports the ESS, promoting harmonisation and standards for statistical production.

CONCEPT	DEFINITION
Data cleaning	Data cleaning (or data cleansing) is the process of detecting and removing errors and inconsistencies from a source of data. It is usually done through the application of checks to identify and delete invalid data. Data cleaning can be considered as a form of data editing including, in particular, the application of validity (e.g. syntactic) and consistency (e.g. semantic) edits.
Experimental statistics vs regular production of statistics	Experimental statistics with mobile phone data refer to the exploration and analysis of data derived from mobile devices for statistical purposes. This may involve utilising data collected from MNOs, mobile applications, or other sources. Experimental statistics often involve innovative methodologies and techniques to derive insights and uncover patterns from these datasets. They are typically conducted in a controlled environment, with a focus on exploring new data sources, testing hypotheses, and developing proof-of-concept studies. Regular official statistics production based on MNO data refers to the integration of MNO data into the official statistical systems of NSIs or other authoritative organisations. This involves establishing robust methodologies, data governance frameworks, and quality assurance processes to ensure the reliability, accuracy, and relevance of the statistics generated. Regular official statistics production entails using MNO data as an input source alongside other traditional data sources. These statistics undergo rigorous validation, analysis, and quality control to meet the standards and requirements set by national statistical systems (NSS) and international statistical frameworks. The objective is to provide reliable and official statistical information that is widely recognised and used for decision-making, policy formulation, and research. This project will provide a standard for the regular production of statistics based on MNO data exclusively (e.g. data from mobile applications other than those of the MNO is excluded)
Functional module	A functional module refers to a modular and self-contained component that performs a specific task or (a set of) functions within the pipeline. Each functional module serves as a fundamental unit that contributes to the overall functionality and workflow of the pipeline. Functional modules are designed to handle a particular step or stage in the data processing pipeline. They may perform (a set of) functions such as data ingestion, data transformation, data validation, statistical calculations, data aggregation, and output generation. These functional modules are interconnected and work together to form a cohesive and automated process for processing and analysing data.
Label	In this project we use the term 'label' to identify a particular characteristic of a statistical unit. In addition, we distinguish the operation of 'creating a label' (i.e. mark with a label, classify the unit in a particular category or classification) from the operation of 'using a label' (i.e. that is simply reading the label that has been already affixed to the unit).

CONCEPT	DEFINITION
Market Share	The percentage of total subscribers that a particular MNO controls within a specific country or a geographic region.
MNO apps	An app, short for 'application', refers to a software program designed specifically for mobile devices such as smartphones or tablets. By MNO app we are referring to an app provided by a mobile operator. With this app, users can conveniently access and manage their account-related information. They can view details such as their billing statements or account balances, gaining insights into their mobile usage and expenses. Furthermore, this app also generates GPS data, capturing location information from the user's device. In certain cases, these GPS data can be utilised as a correction set or for validation purposes within a particular method or process. The app collects and provides this data, enabling enhanced functionality or ensuring the accuracy and integrity of the associated procedures.
MNO data	The term MNO data refers to the set of different types of data provided by MNOs. This includes MNO Event Data and MNO Network Topology Data.
MNO Event Data	MNO Event Data may include: • CDRs and IPDRs • Signalling data • GPS data collected though MNO apps
MNO Network Topology Data	MNO Network Topology refers to the physical layout and configuration of the mobile telecommunications network. MNO Network Topology Data include the cell_id and coverage areas of the network cells or, alternatively, data that allows the estimation of such coverage areas, such as the location of the antennas, their orientation, characteristics (e.g., the technology: 3G, 4G, 5G, etc.), emission power, etc.
Pipeline / Data processing flow	In software engineering, a pipeline refers to a structured sequence of steps or stages through which software code (organised in modules or blocks) processes the data flows for producing certain outputs (n.b. in the case of this project, these outputs are different types of statistics). It is a systematic approach to process the data.
Population	**Population** refers to the total membership of a defined class of people, objects or events. The 'Population' is used to describe the total membership of a group of people, objects or events based on characteristics, e.g. time and geographic boundaries. Examples: • Adult persons in Germany on 13 November 1956 • Computer companies in Spain at the end of 2012 • Universities in Italy on 1st January 2011 Source: Clickable GSIM v1.2 - Clickable GSIM v1.2 - UNECE Statswiki The **Target population** is the universe about which information is wanted and estimates are required. The Target population is the set of the Statistical units.

CONCEPT	DEFINITION
	Source: [ESSnet on quality of multisource statistics - KOMUSO	CROS (europa.eu)](#)
Relevance (quality aspects)	Relevance is an attribute of statistics measuring the degree to which statistics meet current and potential needs of the users. (Source: [KS-GQ-19-006-EN-N.pdf (europa.eu)](#))
Quality in statistics	Quality is the degree to which a set of inherent characteristics of an object fulfil requirements. In the context of the ESS: • The object may be a statistical output, service, process, system, methodology, organisation, resource, or input. • Characteristic means a distinguishing feature. • Inherent means existing in the object, not assigned to it. • Requirement means a need or expectation that is stated, generally implied, or obligatory. Source: [European Statistical System (ESS) Handbook for Quality and Metadata Reports — re-edition 2021](#)
Reference metadata / Explanatory metadata	Reference metadata are metadata describing the contents and the quality of statistical data. Reference metadata are also called explanatory metadata. They include explanatory texts on the context of the statistical data, methodologies for data collection and data aggregation as well as quality and dissemination characteristics. Reference metadata include: • 'conceptual' metadata, describing the concepts used and their practical implementation, allowing users to understand what the statistics are measuring and, thus, their fitness for use; • 'methodological' metadata, describing the methods used for the generation of the data; for example, sample design, collection methods, editing processes; • 'quality' metadata, describing the different quality dimensions of the resulting statistics; for example, timeliness, accuracy. Reference metadata do not define the actual structure of a dataset (structural metadata do this). In the ESS, the Single Integrated Metadata Structure (SIMS) is the standard for presenting reference metadata. Sources: [European Statistical System (ESS) Handbook for Quality and Metadata Reports — re-edition 2021](#) [SDMX Glossary version 2.0, October 2018)](#)
Reference scenario	The project aims to develop the methodological and quality framework and methods for a reference scenario. The term reference scenario' refers to the set of assumptions about the context in which the proposed pipeline will be used by NSIs. The reference scenario foresees the integration of data from multiple MNOs and that a single statistical office has the possibility to reuse (directly or indirectly) the data of all the main MNOs in each EU country, and that final statistics are produced for the integration of all principal MNOs.

CONCEPT	DEFINITION
	Note: The project will develop a 'demonstrator scenario' as a scaled-down version of the reference scenario, with the necessary modifications to deal with current limitations for the testing on real data from participating MNOs (e.g. specific SDC related measures in each national context).
Roamer	SIM or mobile device which seeks or uses an MNO service in a geographic area outside the area served by the MNO with whom it is registered. According to the MNOs and the corresponding Countries involved, we can then distinguish between 'inbound', 'outbound' and 'domestic roamer': Inbound/outbound refer to a roamer from/to a different Country, while 'domestic roamer' refers to a roamer within the same Country.
Signalling data	Signalling data are generated by monitoring signalling traffic in support of network operations. Events are created regularly throughout the day, even if the subscriber does not use any service, which typically increases the number of records per subscriber with respect to other event-based source data, such as CDRs and IPDRs. In other words, while CDRs are triggered by events associated to human activities (e.g. making or receiving a phone call or SMS, starting a data connection), signalling events are triggered automatically by the mobile device or by the network and are only loosely coupled with user activity. Signalling data are more voluminous, frequent and informative, but also more complex to extract and to process than CDRs. Signalling data are, in general, the preferred option for the purpose of analysing population presence and mobility. Source: *(adaptation from)* [Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System (europa.eu)](#)
Spatial/Temporal resolution	Resolution refers to the ability of an instrument or a methodology to separate two points in space or two events in time. The **spatial resolution** is the minimum distance in space between two points that the instrument or the methodology can distinguish. Similarly, the **temporal resolution** is the minimum time difference between two events that the instrument or the methodology can distinguish. Two events occurring at a time distance below the temporal resolution are treated like one. Two points at a spatial distance below the spatial resolution are treated like one. In the processing pipeline, the spatial and temporal resolution of the outputs are use case-specific and can be different from the resolution used along the data processing flow.
Statistical indicator / Indicator	A statistical indicator is the representation of statistical data for a specified time, place or any other relevant characteristics, corrected for at least one dimension (usually size) so as to allow for meaningful comparisons. It is a summary measure related to a key issue or phenomenon and derived from a series of observed facts.

CONCEPT	DEFINITION
	Source: Glossary:Statistical indicator - Statistics Explained (europa.eu)
	The Multi-MNO project introduces several use cases that involve the processing of MNO data and lists for each use case the statistical indicators that can be produced. In the reports from the project, the terms statistical indicator and indicator are used interchangeably.
Structural metadata	Structural metadata are metadata that identify and describe data and reference metadata. Structural metadata are needed and used to identify, formally describe or retrieve statistical data, such as dimension names, variable names, dictionaries, dataset technical descriptions, dataset locations, keywords for finding data etc. For example, structural metadata includes the titles of the variables and dimensions of statistical datasets, as well as the units employed, code lists (e.g. for territorial coding), data formats, potential value ranges, time dimensions, value ranges of flags, classifications used, etc.
	Structural metadata are needed to identify, use, and process data matrixes and data cubes, including. names of columns or dimensions of statistical cubes.
	Sources:
	European Statistical System (ESS) Handbook for Quality and Metadata Reports — re-edition 2021
	SDMX Glossary version 2.0, October 2018)
	Eurostat's Concepts and Definitions Database - CODED: General Statistical Terminology
Statistical Unit	A Statistical unit is the unit of observation or measurement for which data are collected or derived.
	Source: ESSnet on quality of multisource statistics - KOMUSO | CROS (europa.eu)
Tag	Used as synonyms of Label
Testing	Testing involves creating specific test cases and datasets, including both typical and edge cases, to assess the software's behaviour and identify any potential issues or errors. Testing tasks aim to simulate real-world data and processing scenarios that the software is expected to handle. This includes scenarios with large datasets, outliers, missing values, different data formats, and variations in input parameters. By executing these testing tasks, developers can evaluate the system's performance, identify and fix any bugs or inconsistencies, and ensure the software functions as intended under different conditions.
	Note: In the context of this project, the testing/testing tasks are different from the testing of the 'demonstrator scenario' and these terms shall not be used interchangeably.
Traveller	Someone who moves between different geographic locations, for any purpose and any duration.
	Source:

CONCEPT	DEFINITION
	Eurostat(2014) Methodological manual for tourism statistics — 2014, https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-14-013
Tourist	A visitor whose trip includes an overnight stay Source: Eurostat(2014) Methodological manual for tourism statistics — 2014, https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-14-013
Use case	In general, a use case is defined as a specific situation in which a product or service could potentially be used. In the Multi-MNO project, use cases are application scenarios in which MNO data can be used to obtain the different statistical products/outputs through the high-level methodology framework. In D2 report Volume II, the use cases are described according to a 'conceptual' metadata approach/template, which defines the concepts used and their practical implementation. IN D2 report Volume III, the conceptual description of the statistical products based on MNO data is complemented with the concrete specific methodology details, i.e. full methods description for the generation of the data that is used for the calculation of the statistical indicators proposed.
User_ID	User_ID refers to a unique identifier assigned to an individual user or device. It is a string of characters or numbers that helps to distinguish and associate the signalling events to a specific user within the mobile phone data, usually a hash from the IMSI.
Visitor	Traveller taking a trip to a main destination outside his/her usual environment, for less than a year, for any main purpose (business, leisure or other personal purpose) other than to be employed by a resident entity in the country or place visited. These trips taken by visitors qualify as tourism trips. Source: Eurostat(2014) Methodological manual for tourism statistics — 2014, https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-14-013