|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | Logo STATEC - Institut National de la Statistique et des Etudes Economiques | Logo von Statistik Austria |
|  |  |  |  |  |  |



**ESSnet Smart Surveys**

**Grant Agreement Number: 899365 - 2019-DE-SmartStat**

<https://webgate.ec.europa.eu/fpfis/wikis/display/EstatBigData/ESSnet+Smart+Surveys+2020-2021>

**Workpackage 3**

**Development of a conceptual framework, reference architecture and technical specifications for the European platform for Trusted Smart Surveys**

**Deliverable 3.4 Final Report**

**29-06-2022**

**Prepared by:**

Mauro Bruno, Massimo De Cubellis, Fabrizio De Fausti, Claudia De Vitiis, Francesca Inglese, Giuseppina Ruocco, Monica Scannapieco (ISTAT, Italy)

Nils Meise (DESTATIS, Germany)

Jacek Maślankowski (GUS, Poland)

Matjaz Jug, Joeri van Etten, Rob Warmerdam (CBS, The Netherlands)

Franck Cotton (INSEE, France)

ESSnet co-ordinator:

Work package Leader:

Claudia De Vitiis (ISTAT, Italy)

[devitiis@istat.it](mailto:devitiis@istat.it); +390647737401

Table of contents

1. Introduction - WP3 general objectives and activities
2. Summary of deliverable 3.1 – The preliminary framework
3. Summary of deliverable 3.2 – Report on the Proofs-of-Concept
4. Summary of deliverable 3.3 – The improved framework
5. Final remarks: main findings and key messages

# Introduction - WP3 general objectives and activities

The term “smart surveys” refers to surveys that use smart personal devices, equipped with sensors and mobile applications, to collect data. The concept of smart surveys goes well beyond the mere use of web-based (online) data collection that essentially transforms the paper questionnaire into an electronic version. Smart surveys involve dynamic and continuous interaction with the respondents and with their personal device(s). They combine data collection modes based on input from the data subjects (active data) with data collected passively by the device sensors (e.g. accelerometer, GPS, microphone, camera, etc.).

The term “Trusted Smart Surveys” (TSSu) refers to an augmentation of smart surveys by technological solutions that collectively increase their degree of trustworthiness and hence acceptance by the citizens. Constituent elements of a trusted smart survey are the strong protection of personal data based on privacy-preserving computation solutions, full transparency and auditability of processing algorithms. TSSu may also combine smart and traditional data sources. Regardless the type of smart data considered, these new scenarios entail the revision of the statistical process to guarantee accuracy and reliability, according to the principle of accountability and transparency of Official Statistics.

The overall goal of the ESSNet Trusted Smart Surveys (TSSu) was twofold: to define the specifications for the European platform supporting the use of shared smart survey solutions and to assess the usage of applications for European social surveys, such as the Time Use Survey (TUS) or the Household Budget Survey (HBS) (this aim was the objectif of the Work Package 2).

In this context, the Work Package 3 (WP3) carried out preparatory work to design a European-wide platform, to share and re-use smart survey solutions and components for building the TSSu platform. The main goal of Work Package 3 was the development of a conceptual framework, a reference architecture and technical specifications for a European platform for trusted smart surveys, including support for secure private computing to avoid data concentration (e.g., secure multiparty computation), full transparency and public auditability, also including a configurable set of individualised incentive schemes. The work should include any methodological issues related e.g. to quality framework and quality standards, developments related to other standards such as GSBPM (Generic Statistical Business Process Model) and GSIM (the Generic Statistical Information Model), metadata, business enterprise architecture, IT infrastructure, etc.

As stated in the grant agreement, the initial idea of the platform should fulfil the following requirements:

1. Flexible European platform for trusted smart surveys: the platform should be implemented as a set of common (horizontal) functions and configurable services that can be used to build particular instances of TSSu for specific application domains and/or target areas.
2. Development at the European level: the platform should be used independently by NSIs to perform national surveys but could also serve as basis to launch European transnational surveys.
3. Modular, evolvable, extensible and agnostic to particular application domains. It should provide ready-to-use solutions for horizontal functions. It should allow each platform user (i.e., any ESS member) to instantiate a specific trusted smart survey by selecting and configuring different modules.
4. Support for secure private computing to avoid data concentration (e.g., secure multiparty computation).
5. Address methodological issues. New methodological challenges rise in the context of TSSu, e.g. analysis of new data sources, analysis of sensor data, secure private computation. The platform should offer statistical services to address such issues.

The goals of WP3 were very challenging and ambitious, defined at a very high level: we were asked to describe the methodological, architectural and technological characteristics of a generalized and domain agnostic platform. We did this work from a top down perspective, which at least initially ignored specific aspects linked to specific survey contexts. The initial idea for the platform was in accordance with the Eurostat objectives.

The first line of activity consisted in the conceptualisation of a general framework for trusted smart surveys, based on a top-down approach. The design of a reference architecture has started from existing frameworks, up to detail technical specifications. The following aspects were addressed among others:

* ontologies of concepts enabling knowledge sharing and development of semantics;
* measurement methods;
* smart survey methodology;
* technical infrastructure;
* privacy preservation, confidentiality protection, data and process governance;
* individualised incentives and effective communication;
* definition of required metadata throughout the process.

The second line of work consisted in the development of Proofs-of-Concept, in the form of modular prototype elements for essential aspects of the architecture such as:

* active and passive data collection;
* the use of machine learning (ML) for the identification of activities, , missing data, etc.;
* privacy-preserving computation solutions;
* full transparency and auditability of processing algorithms;
* integrating of incentive schemes into the platform;
* front-ends for configuration (allowing survey managers to instantiate and run new surveys with full support for multilingual needs).

The activities to pursue these objectives were organized in two main tasks, structured into sub-tasks, each one coordinated by a responsible partner, as shown in the following tables:

*3.1 Framework Conceptualisation and development of a general platform for trusted smart surveys*

|  |  |  |
| --- | --- | --- |
| 3.1.1 | Smart survey methodology | Destatis Istat Insee |
| 3.1.2 | Technical infrastructure | Gus Cbs Destatis Istat |
| 3.1.3 | Integration in existing architectural frameworks | Istat Insee |
| 3.1.4 | Preservation of privacy and transparency | Cbs Istat |
| 3.1.5 | Incentive schemes | Destatis Istat |
| 3.1.6 | Metadata - Process Auditability | Insee Cbs Istat |

*3.2 Development of proofs-of-concept*

|  |  |  |
| --- | --- | --- |
| 3.2.1 | Data collection and survey methodology | **Istat** Destatis Cbs |
| 3.2.2 | Use of machine learning for evaluating collected data | **Istat** Destatis Cbs |
| 3.2.3 | Privacy-preserving computation solutions | **Cbs** Istat Insee |
| 3.2.4 | Process auditability solutions | **Insee** Istat |
| 3.2.5 | Integrating of incentive schemes into the platform | **Destatis** |

The activities of the Work Package 3 carried out in year 2020, aimed at the definition of a preliminary framework for the TSSu platform. During this period, the activities followed the structure of the first task and the division into sub-tasks, each one coordinated by a responsible partner. The main goal of the work carried out in 2020 was the deliverable D.3.1 “Report on the Preliminary Framework”, produced in February 2021.

The second phase, which began in autumn 2020, consisted in the development of Proofs-of-Concept, in the form of modular prototype elements for essential aspects of the architecture such as:

* active and passive data collection and the use of machine learning for the identification of activities, identification of parallel tasks , missing data, etc.;
* privacy-preserving computation solutions;
* full transparency and auditability of processing algorithms;
* integration of incentive schemes into the platform.

The planning of the PoCs was described in Deliverable 3.1 where the experimental activities were grouped into two sets, according to the perspective of the issues addressed: one group with a methodological point of view, while the second subset concerning more technical and architectural aspects. Furthermore, this distinction follows the structure of the two working groups set up within ESSNet with the dual purpose of linking the activity of Work Package 2 and Work Package 3, and linking the activities of the sub-tasks within WP3.

In the deliverable 3.3 the framework took a step towards the definition of the platform specifications and requirements; the preliminary analysis of the 3.1, the results of the PoCs and the inputs from WP2 were used to model the architecture and the infrastructure described in terms of:

* Improved framework: description of TSSu platform components at conceptual level, in terms of business functions and application components.
* User stories: improved framework at work, putting together the most challenging components, showing how they can be deployed in specific cases.

For the description of the structure and the components of the platform relevant inputs came from the pilots carried out by Work Package 2, in particular from the functional and technical description of the pilots together with the (WP2), described in deliverable 2.2 and 2.5 on modularity and shareability respectively.

In the following of this final deliverable 3.4, summaries of deliverables 3.1, 3.2 and 3.3 are reported (chapters 2, 3 and 4), while the final section contains closing remarks and key messages coming from the results of work-package 3 activities. The bibliographical references can be found in the previous deliverables.

**Interaction with WP2 and working Groups**

From the beginning of the project, a gap between the two work-packages was evident and an effort was made to establish an interaction and a bridging between the different points of view. In the continuous discussion with WP2, the different perspectives emerged, deriving mainly from the opposite starting points, empirical for WP2 and theoretical for WP3.

The experience and the know-how of NSIs and researchers who have been experimented smart surveys have been essential inputs for the definition of the framework / platform. To improve the discussion and the exchange of different perspectives and experiences two Working Groups were established (Methodological issues and Technical issues), numerous meetings were organized on specific topics, putting together different points of view and skills.

From the interaction and discussion with WP2, WP3 gradually realized that absolute generalization was not feasible, although some generalized smart data pipelines could be defined in the platform, depending on sensor and data types. In the discussion with WP2 members, which have a more practical approach than WP3 perspective, a common agreement was that initially the platform could be a repository of software solutions, developed with the aim of being general and domain agnostic, but aware that ad-hoc platform components must be implemented to deal with specific issues.

As the pandemic led to the postponement of the third phase of the HBS pilot and the cancellation of the third phase of TUS, it was possible to partially use the data of the Wp2 pilots for the PoC experimental phase. In fact, only the data from the WP2.3 pilot on Health conditions were used. In order to have another data set to experiment the ML procedures, data provided by the University of Trento was obtained, collected through the ILog app (see deliverable 3.2).

***Working Group Methodology***

The working group methodology had two main objectives. First, to close the gap between the high-level domain agnostic methodological description of the platform (WP3) and the hands-on methodological approach of the pilots (WP2). And second, to support the PoCs (WP3) and pilots (WP2) with additional insights. In 2021 the working group had six meetings that evolved around common issues in WP2 and WP3. The agenda of the meetings included the following topics.

* Exchange on available data and data collection in WP2, which included data donations for the use in WP3 PoCs;
* Discussions of the smart data lifecycle, data processing and data analysis (with an emphasis on machine learning);
* Discussions on smart survey error with lessons learned from the WP2 pilots and how contextual data could be used for mitigation. In particular, how the use of sensors and machine learning has an impact on error treatment;
* Exchange on paradata collection during WP2 field tests and lessons learned.

The meetings build several bridges between the different scopes of the working packages by establishing a discussion on the different approaches, common solutions and especially hands-on problems with data and error treatment. In particular, these discussions helped to understand what are relevant problems that the design for the European platform needs to cover, requirements for monitoring and metadata management, and the importance of human intervention or the human in the loop. The latter was of significance to understand what parts of TSSu could be automated and for which parts error treatment based on (human) domain knowledge is crucial.

***Working Group on Architecture***

The working group was established with the following (main) objectives.

* Closing the gap between the high-level description of the TSS platform components (WP3) and the hands-on technical implementation of the pilots (WP2). More precisely, for each dimension of analysis (methodological, technological and metadata), the group worked to foster the alignment between WP2 pilots and WP3 proof-of-concepts.
* Supporting the PoCs (WP3) and pilots (WP2) with additional insights. The results of the pilots have been used to test the theoretical models defined in WP3.

In 2021 the working group had several meetings that evolved around common issues in WP2 and WP3, allowing to achieve a common vision on the global outcomes of the ESSNet project.. The agenda of the meetings included the following topics, regarding the Platform Architecture, with the main goal of providing a set of technical requirements for the TSS platform, based on WP2 field experience.

* Frontend: the platform should provide a data collection tool that allows to design questionnaires that run on mobile devices, web browsers, etc. Further, the platform should offer applications that allow to design surveys (load samples, assign questionnaires to enumerators, etc), monitor fieldwork activities, run data processing algorithms, and analyse statistical outputs (data analytics, business intelligence)
* Backend: the backend should be designed and implemented according to modern cloud-native architectures, i.e., loosely coupled microservices that execute a different function, whether implementing capabilities, communicating, or running processes. The infrastructure hosting the backend could be on-premises, in the cloud (Amazon, Google, Azure) or both.
* Data storage: data should be stored in Relational databases (MySQL, PostgreSQL, etc) in NoSQL databases (MongoDB, Redis, Couch DB, etc) or both? Where should data be stored (on-premises, at national level, in the cloud)? We can imagine different data models, e.g. microdata stored in NoSQL format, metadata stored in Relational database.
* Privacy issues: privacy by design should be a guiding principle in the design and implementation of the platform. A key aspect to be addressed in the platform is related to input-privacy preserving techniques. The group worked both on the theoretical aspects and on the software/hardware requirements related to input privacy.

The members of the group (and the invited experts) shared experiences on architecture modelling and implementation activities, such discussions allowed to gain a deeper insight on the architecture problems involved in the design of a TSS platform managed at European level.

WP3 deliverables can be found at

<https://ec.europa.eu/eurostat/cros/content/essnet-smart-surveys_en>

# Summary of Deliverable 3.1- The preliminary framework for the Smart Surveys

**Table of contents**

|  |  |  |
| --- | --- | --- |
| **Activities in Work Package 3** | **Chapter of the deliverable** | **Leading and participating beneficiaries** |
| 3.1 Framework | Introduction: WP3 overview and general features of the platform | **Istat** |
| 3.1.1 Smart survey methodology | 1. Smart Survey Methodology | **Destatis** Istat Insee |
| 3.1.2 Technical infrastructure | 2. Technical infrastructure | **Gus** CBS Destatis Istat |
| 3.1.3 Integration in existing architectural frameworks | 3. Integration in existing architectural frameworks | **Istat** Insee |
| 3.1.4 Preservation of privacy and transparency | 4. Preservation of privacy and transparency | **CBS** Istat |
| 3.1.5 Incentive schemes | 5. Incentive schemes | **Destatis** Istat |
| 3.1.6 Metadata – Process Auditability | 6. Metadata – Process Auditability | **Insee** Cbs Istat |
| Planning of Proof-of Concept (task 3.2) | ANNEX – Planning of Proof-of Concept | **Istat** CBS Destatis Gus Insee |

The deliverable represents the product of the first year of activity of Work Package 3 on the conceptual framework for Trusted Smart Surveys. The framework is here in a preliminary stage; all the sub-activities have started from a very high level following a top-down approach. This deliverable addresses the main statistical and technological aspects involved in planning and conducting a smart survey, focusing in particular on those steps deserving a different specification with respect to traditional surveys.

The deliverable is divided into six chapters, corresponding to the six sub-tasks 3.1.1-3.1.6 and include an Annex outlining the Proof-of-Concept.

**Task 3.1.1 Smart Survey Methodology**

The activity of task 3.1.1 aims at developing a robust smart survey methodology. It explores design requirements for TSSu in contrast to traditional paper-based or online surveys. The main problems addressed here are related to: sensor data from a variety of devices that are not standardized in structure, format or availability; innovative ways of handling sensor data (including ML algorithms); sources of error in TSSu and error management; GSBPM phases involved in TSSu; automation of smart surveys and needed staff profiles.

This chapter consists of three sections:

* Smart data features.
* Errors in smart surveys.
* Smart survey processes.

The first section focusses on the main methodological aspects and features of TSSu, designs and new data collection methods and presents an overview on: mobile devices and wearables with detailed information on the sensors and data they provide; ML algorithms and issues related to the quality of (sensor) data, such as training models and metric to evaluate ML algorithms.

The second section analyses the sources of error in TSSu considering the technologies and instruments used to collect data and the new type of data sources (sensor data acquired actively and/or passively through apps). The purpose of this analysis is to highlight all the elements necessary for a redefinition of the Total Survey Error (TSE) framework for smart surveys which considers the new sources of error, both in representation (selection) and measurement (traditional and sensor data). For analysing data accuracy in TSSu, the TSE for traditional surveys and the Total Error (TE) defined for big data are considered as reference frameworks: the last one to understand the sources and nature of new types of error in sensor data acquisition and management.

The final section outlines a preliminary methodological framework for TSSu taking GSBPM as a reference statistical model to facilitate connections with the architectural system, in particular with data collection and data processing components of the platform. The focus is on the phases and sub-processes that need changes for the use of smart technologies and smart data, innovations that require a new data collection channel, new processing tools, and new skills for managing sensor data. A GSBPM design phase for TSSu is introduced, the take on data collection for TSSu is linked to the GSBPM collect phase and indicators for setup and monitoring are discussed. Moreover, the challenges of processing smart data are expanded by the use of machine learning by drawing a link to the GSBPM process phase. Finally, automation of smart data processing and infrastructure and staff profile needs are discussed.

**Task 3.1.2 - Technical Infrastructure**

Regarding Technical Infrastructure issues, an architecture of two different smart surveys platforms is shown, i.e., HBS (Household Budged Survey) and MOTUS (Modular Online Time Use Survey). Its technical infrastructure analysis was also used to create a generic infrastructure for smart surveys. In this chapter, there is also a list of requirements that should be fulfilled by the smart survey application. It is important to note that these two smart surveys applications are still under development.

This chapter focuses on the analysis of the as-is architecture of HBS (Household Budget Survey) and MOTUS (Modular Online Time Use Survey) platform, as of 2020. The current work done is a preliminary step to develop a to-be architecture in the next phase of smart surveys maturity. The next step would be to develop a Proof-of-Concept for the smart surveys.

The chapter is divided into three parts where they are discussed:

* A general information on possible ways of data collection through the mobile devices was presented.
* The requirement analysis for smart surveys, i.e., mobile applications used to collect statistical data.
* The logical components of technical architecture according to the BREAL - Big Data REference Architecture and Layers Application and Information Architecture.

The current applications of smart surveys allow to collect selected data without user interaction. However, most of the data must be collected in traditional way, i.e., by the respondent input. It shows that it is not possible to gather all information just based on sensors. Such data that cannot be gathered through the sensors includes, e.g., type of activity at home (e.g., reading books, watching TV, cooking) for Time Use Surveys. In Household Budget Survey it is necessary to write specific goods bought if there is no information on the bill or it is not possible to take the photo of the products. Moreover, sometimes the OCR (Optical Character Recognition) to change the bill into text may be not sufficient to detect a specific product, as the name on the bill may be not precise (e.g., producer name instead of the product name).

The analysis of logical components used in the smart surveys shows that the main function of smart surveys is to support traditional surveys. It will not be possible to replace the traditional questionnaires with data based on sensors. Thus, the hybrid application that can collect selected data with sensors and other, missing data with electronic questionnaire can be the model of the smart surveys in most of the statistical domains.

**Task 3.1.3 – Integration in Existing Architectural Frameworks**

The main goal of task 3.1.3 is to provide an overview of reference standards and initiatives preceding TSSu exploration. The main goal of this analysis is the alignment of WP3 activities with existing architectural frameworks, to benefit from the achievements of related experiences. The alignment with existing frameworks is also compliant with TSSu driving principles addressing the privacy issue, the new methods for smart data processing, as well as process transparency, accountability and smart data life cycle.

Starting from an overview of official statistical standards and legal framework, the chapter focuses on the main results of the ESSNet Big data II, especially on the Big Data REference Architecture and Layers (BREAL). A brief description of relevant data platform projects complements the introductory analysis. Finally, the last section analyses the key elements to consider for modelling the business layer of a TSSu platform. The activities of task 3.1.3 have delivered preparatory work to share and re-use smart survey solutions and components for the development of the TSSu platform. This platform, conceived as a set of common (horizontal) functions and configurable services, will allow building instances of TSSu for specific domains and/or target areas.

**Task 3.1.4 – Preservation of privacy and transparency**

NSI's should look for ways to enable leveraging new data-sources, while preventing privacy being breached, so the aim of task 3.1.4 is to assist in this process.

This part of the deliverable provides a set of tools that might help with understanding and dealing with the privacy considerations involved in the aforementioned new forms of data collection. In all data collection there is an inherent concern regarding privacy, with varying levels of possible concern. Much research has been, and is being done to enable data mining while ensuring at least some level of privacy. In the sharing of data, the objective would thus be to not only store, but also collect and process these in a privacy preserving manner.

The chapter consists of three sections.

The first section presents an overview of the Privacy-by-design framework for handling private information. This framework is subsequently converted into a set of guidelines, which should assist in process of making correct decisions concerning how private data should be dealt with by an application which collects such data.

The second section is dedicated to techniques for preserving privacy while maximizing data utility. An overview of a number of such techniques (e.g. differential privacy, homomorphic encryption, secure multi-party computation, etc.) is given. The benefits and drawbacks are highlighted, as well as typical use-cases. Finally, a discourse on what these techniques could mean TSS applications is presented. In addition, a number of scenarios in which these techniques could be applied in the smart survey process are described.

In the final section, the issue of user authentication is tackled. Naturally, the identity of TSS participants should be verified. However, this should be done using a minimal amount of information to reduce privacy impact. A protocol for user authentication using a minimal set of characteristics is therefore presented in the form of IRMA, short for 'I Reveal My Attributes'. Finally, an example implementation of IRMA in the context of a TSS application is given.

The development of smart surveys should follow the privacy-by-design principles, that applied in the context of the TSSu, leads to six guidelines that should be considered. Following these guidelines would make sure that respondents remain in control of their data, have a clear understanding of what their data is used for and allow them to selectively change which of their data is shared. Guidelines for the smart survey context could be developed by:

1. Using solid, existing privacy guidelines.
2. Specifying them to assist the users’ understanding.
3. Assessing both privacy risks, and research value, aiding in informed decisions between trade-offs.
4. Suggesting methods and technologies to control processes based on the guidelines and assessments.
5. Providing a worked through example by looking at a mobility application for a household study.
6. Serving as a handbook for reference usage and further developing in the context of collecting and processing (sensor) data for NSI’s, specifically for using in smart surveys.

**Task 3.1.5 – Incentives Schemes**

The activity of Task 3.1.5 aimed at exploring incentives for TSSu. Manual data entries can be cumbersome for respondents and potentially lead to missing entries and drop outs. This activity investigates how gamification could be used as a strong incentive scheme to engage respondents more in the data collection process.

Ideally, incentives make surveys a fun and engaging activity, thus reducing the burden for respondents. However, it should not turn into a huge burden on the side of the NSI, which has to allocate more resources of time and personal for design and during deployment. Even in a full-automated environment, gamified incentives require additional manual steps, e.g. the development of personas or definitions of user types. In addition, personas and user types require constant development for future surveys. A successful implementation also depends on (new) experts in UI or HCI design.

Gamification is a proven tool to foster engagement in a given activity and widely used in different fields of business’ and tested in online surveys and app-based scenarios. Also, respondents are used to gamification elements during leisure activities and increasingly in education and work settings. However, a wide range use of gamification is ethically challenging, because it aims at adapting to socially expected behaviour and rewards conformity.

User types in gamified settings are an orientation for development and allow to adapt for different types of respondents not only in app design, but also in communication in order to appeal to specific audiences. But user types are a simplified abstraction of motivation only. Personas offer a similar variety of types with a more in-depth and focussed attention on specific and specified ideas of users. Implementation of incentives fall into wider methodological considerations, bound to the business logic of the production of official statistics (GSBPM) and big data in official statistics (BREAL) and are a challenge for continuous development, integration, privacy and ethics. Some of these challenges we discussed in regard to particular risks of gamification.

The chapter is divided in the following parts:

* An overview of common incentives schemes and an introduction into gamification.
* How gamification can be applied to surveys and how to tailor it to specific groups of respondents.
* A discussion of implementation and risks of gamification.

**Task 3.1.6 - Metadata**

The term ‘Metadata’ refers to many concepts and dimensions, interconnected and with specific features at the same time. In order to contribute to the TSSU platform design, the main objectives of the Metadata task relate mainly to:

* Developing a common semantic representation of the domain covered by the project;
* Providing recommendations and guidelines regarding the capture of metadata all along the operation lifecycle. The whole set of metadata will cover:
* statistical concepts related to the survey theme and objectives;
* collection instruments, variables and codes;
* statistical methodology used in data processing;
* process and lineage metadata;
* quality metadata.

This list of metadata relates to different dimensions: domain concepts, collection tools, regulations, methodology, output description, process tracking and quality.

The main objective of task 3.1.6 is the design of the TSSU platform metadata component, which is a challenging task that requires an in-depth analysis of several aspects, namely:

* Statistical concepts related to the survey theme and objectives;
* Collection instruments, variables and codes;
* Statistical methodology used in data processing;
* Process and lineage metadata;
* Quality metadata.

Starting from a preliminary analysis of standards and frameworks (i.e., ontologies, BREAL Information Architecture, GSBPM and GSIM) related to metadata concepts, the chapter focuses on the following general aspects:

* Data structures;
* Sensor data;
* Process steps;
* Privacy issues;
* Generic functionalities for reuse in different domains;
* Specific characteristics of national contexts.

The approach adopted to fulfill these general objectives is based on the following principles:

* Rely on standards and align with existing frameworks;
* Use active/passive metadata;
* Use-case-driven approach.

In order to overcome specific domain requirements, the main goal of this task is to model a repository with the minimum set of metadata for process standardization and application components reuse. The implementation stage should focus primarily on the description and integration of sensor and traditional data used in TSSu.

# Summary of Deliverable 3.2 - Report on the Proof-of-Concept

**Table of contents**

|  |  |  |
| --- | --- | --- |
| **Activities in Work Package 3** | **Chapter of the deliverable** | **Leading and participating beneficiaries** |
| 3.2 Development of proofs-of-concept | Introduction: Overview on the Proof-of-Concept | **Istat** |
| 3.2.1- 3.2.2 Data collection and Smart survey methodology - Use of machine learning for evaluating collected data | 1. Treatment of sensor data for Smart Surveys through Machine Learning: a Generalized Module. “PoC on methodologies” | **Istat** Destatis |
| 3.2.3 Privacy-preserving computation solutions | 2. Application of Privacy Enhancing Technologies to the TSSu platform | **CBS** Istat |
| 3.1.4 Full transparency and auditability of processing algorithm | 3. Full transparency and auditability of processing algorithms. “PoC on architecture and metadata” | **Istat** Insee |
| 3.2.5 Integrating of incentive schemes into the platform | 4. Integrating of incentive schemes into the platform | **Destatis** |

The deliverable 3.2 is the report of the second phase of WP3 activities, that consisted in the development of Proofs-of-Concept (PoCs). The PoCs aimed at verifying the feasibility of specifics parts of the conceptual framework descripted in deliverable 3.1. This constitutes a step forward in the direction of a more practical definition of the European platform, because experimenting with the different aspects of a smart survey produced a more concrete vision of what can be achieved and with which problems.

The general purpose of the PoCs is to subject the main aspects of the platform described in the preliminary framework to operational verification, to assess whether the characteristics outlined or modeled are applicable in the practice of smart surveys from different perspectives, and to provide evidences and inputs for the definition of the enhanced framework and the specification of the methodological and technical requirements of the platform.

The deliverable is divided into four chapters dedicated to the description of the:

* PoC on methodology and machine learning, conceived for designing and developing a G*eneralized Machine Learning* *Component* (GMLC) for data provided by the same type of sensor (e.g. accelerometer, gyroscope, thermometer).
* PoC on Application of Privacy Enhancing Technologies (PET) which tests the applicability of various privacy techniques to the Trusted Smart Survey platform, and discusses the practical feasibility of the proposed architecture components.
* PoC on incentive schemes, modelling how an incentive in form of gamification can be included into a Trusted Smart Survey and highlighting which gamified elements are suitable for an overall implementation and how they can be used in conjunction with a targeted focus on respondents.
* PoC on architecture and metadata, to test the assumptions made in the previous tasks (3.1.3 and 3.1.6) of Deliverable 3.1, and model the business and application layers, for the design of a set of application services to be provided by the platform. Furthermore, the PoC describes the architectural aspects and the metadata related to smart data processed in the Machine Learning (ML) PoC.

**Tasks 3.2.1-3.2.2: Treatment of sensor data for Smart Surveys through Machine Learning: a Generalized Module “PoC on methodologies”**

The application of Machine Learning (ML) techniques is a central aspect of smart surveys. The PoC of sub-tasks 3.2.1 and 3.2.2, which aims at the definition and implementation of a generalized ML component, wants to contribute to the development of the European platform.

Machine Learning is an essential technique for sensor data processing and the fact that its implementation is agnostic draws great interest.

The goal of the PoC is to design and develop a G*eneralized Machine Learning* *Component* (GMLC), divided into different software modules that will allow it to be applied to different contexts and survey needs. In addition, sensors do not often measure directly variables of interest and the PoC will explore ways to assist. A machine learning module can be generalized to different surveys if it deals with data provided by the same type of sensor (e.g. accelerometer, gyroscope, thermometer).

The GMLC is developed on a cross-survey component performing multi-class supervised classification tasks, although the extension to regression tasks can be implemented. The generalization of the process is realised considering different modules that perform a specific function in the pipeline. Modules can be interposed into the process in a different order and new modules can be added to fit new surveys.

ActivPAL and iLog datasets constitute a baseline for developing a component in terms of modularity. Both data collected from a diary and sensors by Health pilot (Work Package 2.3) and in SmartUnitn(Two) surveys were used to build a more generalised pipeline. The use of data collected through different devices (wearables for Health pilot and smartphones for SmartUnitn (Two) surveys) have implications on the results. In fact, while for the activPAL use case, the performance of the physical activity classifier is good (accuracy of 87.6%), for the iLog use case, the performance of the “mean of transport” classifier is lower (accuracy of 61.6%).

These differences are due to the good quality of the data collected in a controlled environment, such as the laboratory for the annotation of the labels, and also to the wearable instrument activPAL that measures the acceleration with a good sampling rate, and a systematic collection of the associated acceleration signal. In SmartUnitn(Two) surveys, for some classes, the performance is very high when the associated signal is very characteristic (for example because of the presence of high-frequency vibrations). For other classes, for which the acceleration profile is very similar, the performance is lower. The worse performance of this use case compared to the activPAL can be explained by the difference in the types of activities classified in the two cases, and by the differences in the device position, constant in one case and variable in the other. In fact, signals from smartphones present a great variability of the acceleration profile for the same activity, due to the position in which the device is placed.

**Task 3.2.3 - Application of Privacy Enhancing Technologies to the TSSu platform**

The main objective of the PoC is to test the applicability of various privacy-enhancing technologies (PETs) to the Trusted Smart Survey platform, with the purpose of enabling production of statistics without access to respondent level microdata. The first part of this ‘test’ is theoretical. It aims to answer the question: “Given the supposed possibilities and shortcomings of PETs, how can we design a platform architecture that utilizes these technologies to the best of their abilities?” The second part aims to answer questions regarding the practical feasibility of the proposed architecture components. Given limited resources, we chose to focus on the component regarding statistical analysis using secret sharing based MPC.

The first part of the chapter reports an example of platform architecture that allows for the implementation of the PETs required for production of aggregate statistics without compromising respondent privacy. Other issues/aspects are highlighted: that for a well-designed and well-defined survey, the architecture required for oblivious data-analysis is not much more complicated than one would expect for a platform without; quality control, logging and hence survey maintenance without compromising privacy is difficult; design and development of surveys is incredibly difficult given the restrictions of privacy enhancing technologies. These difficulties are inherent to the goal of applying these technologies and can hence not be overcome by technological advancement.

In the second part, the process of adapting the existing PETs to the statistical use case is described by developing a toy model application that enables aggregate statistical analysis using multi-party computation for a general (smart) survey. This showcase should not be interpreted as a demonstration of how to develop a production-level application for privacy preserving statistical analysis. Rather, it presents some very basic ideas and considerations that can be used in the development of such an application. One finding is that while it seems possible to adapt existing open-source technology to statistical use-cases typical for the TSSu platform, the process is highly involved. As such, proper implementation embedded within the TSSu platform likely requires experts in the fields of statistics, software engineering and cryptography.

**Task 3.2.4 - Full transparency and auditability of processing algorithms. “PoC on architecture and metadata”**

The main goal of this PoC is to test the assumptions and benchmark the results concerning:

* The previous inventory of official statistical standards and ongoing projects, released by Task 3.1.3[[1]](#footnote-1) dealing with the preparatory work for the design of a TSSu platform and integration in existing architectural framework
* The approach adopted by task 3.1.6 Metadata and Process auditability, leading to the design of a

Metadata Repository (MR) composed of several subsets, based on existing frameworks and ontologies.

The chapter is structured in two main parts, describing respectively the architectural analysis and the metadata captured for the Machine Learning (ML) PoC.

In the architectural use case, the main steps of a generic TSSu are modelled. More in detail, starting from the description of the statistical process in terms of GSBPM phases and BREAL business functions, the architectural PoC has modelled the business and application layers, to propose a set of application services to be provided by the platform. The analysis of operational models and architectural scenarios has complemented the theoretical assumptions related to the architectural layers. The specification of the dimensions affecting the deployment of the application services, such as data storage and processing environment or adoption of Privacy preserving techniques is essential to define the requirements of the technical infrastructure of the platform.

The second part deals with the description of process metadata related to the tasks performed by the ML generalized module, developed to process accelerometer data. Starting from the conceptual subsets of the MR, the PoC has allowed describing the process steps executed, to track data transformations, to document and assess the methods applied in each step.

The last paragraph summarizes the main results achieved by the PoCs, the main lessons learnt and some open issues to be addressed, to achieve an enhancement of the reference architecture delivered at the end of the first year of the project. One of the main open issues is to highlight the relationship between the dimensions explored by the single tasks, such as: methods, incentives, privacy preservation, process auditability, assessment of smart data quality, technical requirements.

Concerning the main results achieved, the architectural PoC has demonstrated the feasibility of process standardization, to foster the integration between traditional and smart survey tasks. This analysis supports the specification of the main technical requirements of the TSSu platform, and is essential to plan the development activities. The metadata PoC has confirmed the initial assumptions, based on a central repository for the collection and management of the process metadata related to specific tasks of sensor data processing.

**Task 3.2.5 – Integrating of incentive schemes into the platform**

The activity of task 3.2.5 explores incentives schemes in a Proof of Concept to understand how gamification could be utilized for a gamified (smart) surveys. It asks which steps are required to design a gamified survey by looking at different survey phases, and use patterns to understand these phases as the respondents or player journey respectively. In addition, prospective respondents are handled with an abstract concept of user types. The implementation of gamification results in a trade-off: it eases the hard task of completing a longer running survey for respondents, and at the same time requires more and new forms development by a survey agency. These additional development activities are mainly due to the different mode of app-bases surveys and design requirements that result from that decision.

The chapter consists of the following parts. First, an introduction that gives an insight why gamification was chosen and under what assumptions the PoC was conducted. Second, which assumptions are useful when designing a gamified survey for a specific group of respondents (2. Desired User Types). Third, the actual design of a gamified survey, following design principles for apps and gamified apps in particular (3. User Journey). Finally, a conclusion discusses the results and identifies open issues for further investigation (4. Results).

# Deliverable 3.3 - Report on the improved framework

Deliverable 3.3 focused on the conceptual framework of the European Platform for Smart Surveys, main goal of the activity of Work-package 3 (WP3). The last part of the project aimed at producing an improved framework, revised after the Proof-of-Concepts and in light of the results provided by WP2.

In the enhanced framework, the general architecture of the TSSu platform is described in terms of business and application layer, aligned with existing frameworks and official statistical standards, such as the GSBPM (Generic Statistical Business Process Model) and GSIM (Generic Statistical Information Model). The business layer follows the process and sub-process of GSBPM, to identify and standardize the core activities performed by the National Statistical Institutes (NSIs) for smart data collection and processing (What).

The definition of the main process steps fosters the design of the application layer, providing an overview of the main components to be developed for the definition of a common infrastructure (How). These components are the building blocks of the general platform and they are analysed from the different perspectives: Methodology, Privacy and Metadata.

In particular, the methodological aspects of a smart survey deepened in this deliverable are those more tightly related to the collection and exploiting of sensor data and new data sources. Therefore, all the methods that can be applied analogously in a traditional survey are here out of scope. Moreover, the focus here is mainly on the data collection and data processing phases. The other phases of the survey, in fact, such as the sampling design, the integration of data collected with different modes (if a mixed-mode data collection is adopted) and the estimation phases are not addressed as, at least in this phase, the platform aims to develop and support only the smart features.

It is useful to underline that the platform delineated does not foresee only the objective to develop the application tools for the data collection; the tools can be implemented by NSIs autonomously or can be selected from a list of existing tools, such as those reviewed by the Task Force Innovative Tools and Sources. So, the platform could utilize existing tools, fostering their integration in the survey pipeline.

Moreover, the use of a tool and the realization of a survey requires experimentations, through a test phase. The platform can assist the NSIs to project and conduct these tests.

The enhanced framework tries to cover all the objectives of WP3, those addressed explicitly in the PoCs,and those addressed by WP2 in the pilots, concerning mainly the data collection aspects.

In a more practical perspective, the platform outlined by WP3 follows the idea of providing to European NSI statisticians a set of tools for the execution of functionalities to carry out a smart survey, in a modular and incremental approach, starting from what could be implemented in practice. These functionalities could be deployed locally or centrally and should address primarily the strictly “smart” aspects of the survey, such as: the sensor data collection, the monitoring system related to data collection and to sensor data quality, machine learning algorithms for processing sensor data, the metadata management for guaranteeing the auditability of the process, and privacy and security issues.

The deliverable 3.4 was conceived as a consistent document, without direct reference to the tasks of the work-package, and is organized as follows: the architecture of the framework; the functionalities of the building blocks, focusing on methodological aspects, privacy issues and metadata management; the technical requirements of the components; the possible functioning of the platform, through the description of some use cases. The Annex, reporting some operational scenarios, complements the description of the enhanced framework.

**4.1. The architectural framework for TSSu**

The design of the enhanced framework aims at highlighting the relationships between many different aspects, related to three main dimensions:

* Architectural, concerning the design and development of software solutions for smart data collection and processing.
* Methodological, regarding the available privacy preserving techniques, and the different methods for smart data gathering, validation and processing (e.g.: data collection strategy, design of the user interface, edit checks and data quality, machine learning models).
* Technological, dealing with the design of the technical infrastructure according to the privacy preserving requirements, as well as the interaction between the platform components.

Starting from the alignment with existing frameworks, the architectural Proof-of-Concept has analysed these issues, modelling the main process steps for smart data processing according to the Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM) standards, as well as the Big Data REference Architecture and Layers (BREAL) business functions.

An overview of smart data pipeline, from a National Statistical Institute’s perspective is essential to model the business and application layer of the TSSu platform and assess: (i) the impact of smart data sources on GSBPM phases and process steps; (ii) the process steps which can be standardized and executed in a common infrastructure, to identify shareable and reusable application components; (iii) the tasks that can be executed in the TSSu platform or in the respondent’s device, as well as the tasks performed in-house by NSIs; (iv) the best strategy to harmonize specific national needs and the peculiarities of the stakeholders engaged in a TSSu process.

*Business Layer*

The main goal of modelling the Business layer of a TSSu is to identify and standardize the core activities and behaviours (What) performed by NSIs for smart data collection and processing. The adoption of a top-down approach aims at highlighting the interdependence between the key dimensions listed above. In addition, the definition of the main process steps allows to design the application layer, which provides an overview of the software solutions (How) to develop for a common infrastructure.

The design of the business layer has provided an overview of the main tasks to execute, and has highlighted the relationship between the GSBPM phases and BREAL business functions concerning smart data processing (e.g., Acquisition and Recording, Data Wrangling, Data Representation). While “Metadata management” and “Privacy Preserving techniques and management” are overarching business functions, the sub-processes performed in the different GSBPM phases for smart data have been considered as a specialization of the related business functions.

*Application Layer: main building blocks*

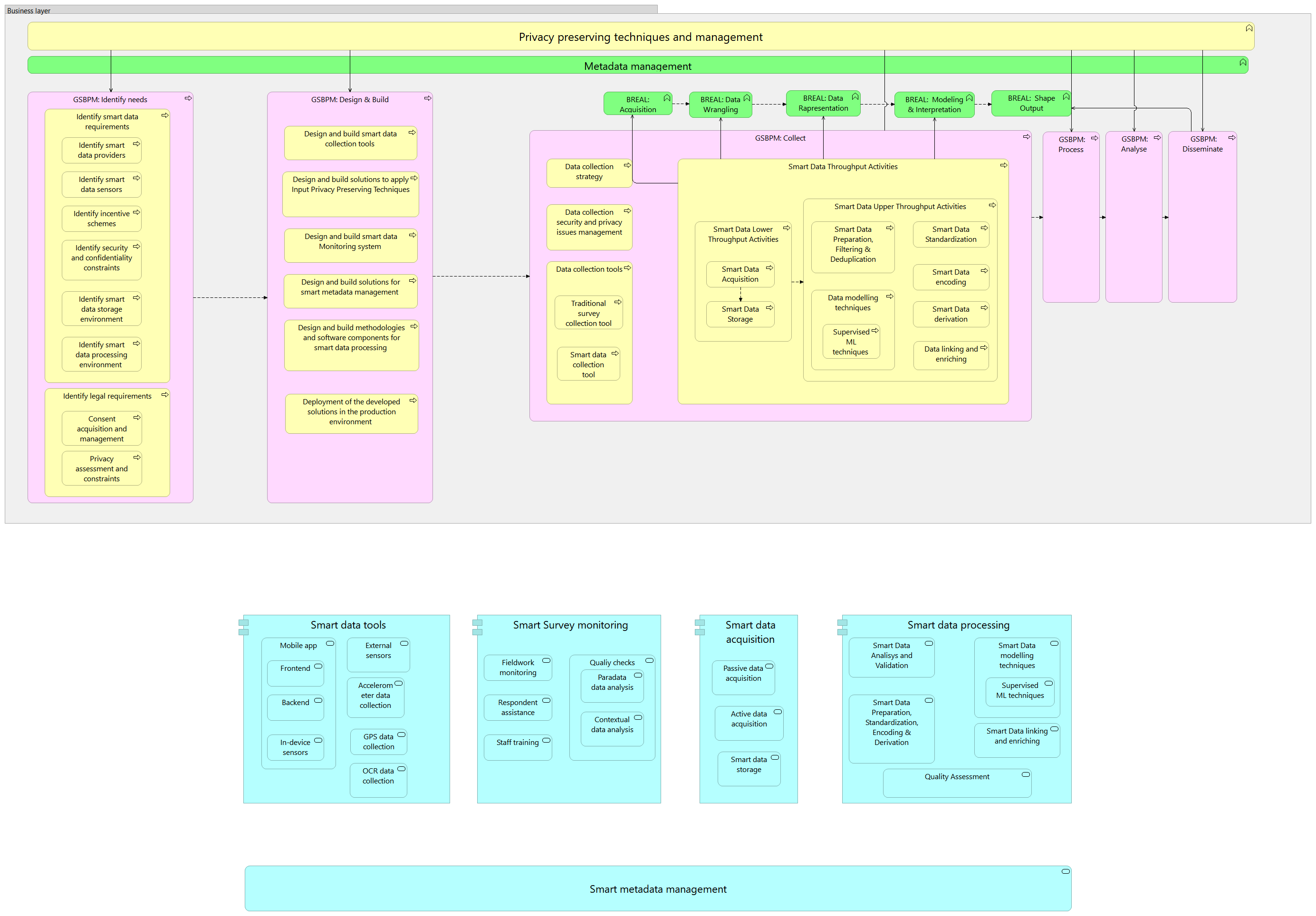
Modelling the business layer is the first stage to define a preliminary set of building blocks needed to run the process steps of a general TSSu. These building blocks are described in terms of functionalities that could be offered by the TSSu platform and will allow to make some assumptions in relation to the service deployment and sharing.

Regarding the quality, security and privacy issues, an analysis of the impact of these dimensions on the building blocks will help to identify the requirements of the technical infrastructure and software solutions. According to the lessons learned in the different WP3 PoCs and in the experiences acquired in WP2 through the pilots, the process pipeline of a TSSu results from the combination of several tasks executed by modular components. The design of data collection and smart data processing building blocks should be driven by:

* The type of data source (private or public data).
* The type of data provider (Respondent, Third parties, public authorities).
* The type of sensors used for data gathering (in-device, or external sensors).
* The storage and processing environment (In-app/NSI/TSSu Platform /Third parties).

In order to design a general pipeline for a TSSu and foster the combination of smart and traditional data sources, the following figure reports an overview of the proposed building blocks and the connections with the metadata building block.

*TSSu Building Blocks*



**4.2. Methodological aspects of TSSu**

The methodological aspects of a smart survey deepened in WP3 activity are those more tightly related to the collection and exploiting of sensor data, collected using specific tools running on a smart device; hence the focus is mainly on the data collection and data processing phases.

WP3 followed the scheme used by WP2 for describing the methodology level, involving the following choices: user interface of frontend, data collection strategy, data quality checks, machine learning models in processing sensor data. In the deliverable, some choices have been explored more in depth by moving on both theoretical and operational aspects.

***Data collection strategy and user interface of frontend***

The data collection strategy includes all smart (and non-smart) data collection tools, their design and the logistics involved. In particular, the building blocks *Smart data tools* and *Smart survey monitoring* in terms of *Respondent assistance* and *Staff training*. This concerns the use of contact modes, contact and reminder strategies, incentive strategies, recruitment materials, use of “non-smart” modes. The data collection strategy for smart surveys utilizes the data collection tools in form of a (mobile) app to provide a frontend for questionnaires or reporting. Further, it provides services to allow passive and active data collection with sensors of the mobile device or to import data from external sensors, which extends the collection range beyond a single device. In addition, questionnaires can be placed in the same app. Thus, an app is a central hub to govern the data collection process. The design goal is to provide a unified survey experience through one standardized frontend.

During data collection, sensors will fail and respondents provide false data by accident or on purpose. A smart survey monitoring mitigates risks by identifying problems early with automated checks and alerts for both the respondent and the survey agency, in order to adapt the data collection strategy and offer further assistance. Such adaptive strategies could entail help for the appropriate use of a device for data collection, e.g. how to position what kind of places or movements to avoid, via tutorials within the app or by offering quick ways to get in touch with trained staff at a help desk.

Further improvements for the data collection strategy include a thorough on-boarding process for the respondents, smart notifications to keep the respondent on track and ultimately incentive strategies like gamification as an integral part of the survey experience.

***Data quality***

The combination of traditional data (provided directly by the respondent using a questionnaire) with big/sensor data (provided in active or passive way) deriving from different sources (internal sensors, external sensors, public online data, personal online data) obliges looking at quality as a requirement that needs the definition of new concepts and metrics and the development of new approaches for the validation analysis.

A general data quality (DQ) framework should be declined right from the data source, the type of data and sensor, but also considering new sources of representation (coverage, participant selectivity, non-willingness to provide sensor data) and measurement (sensor) errors.

Characteristics and properties of sensors data and the quality of measurements must be considered in defining a DQ framework. In fact, the quality of sensor measurements can be affected by limitations of the sensor itself (inaccuracy, time inequivalence), the heterogeneity of devices (device inequivalence), the behavior of the participants in the survey, etc. Different aspects should be considered if sensor data derive from third parties, but in this case the flow of the data acquisition would be different and also the quality would be carried out in a different perspective.

The chapter deals with the problem of data quality for sensors considering both theoretical and practical aspects. The main in-depth parts are:

* The need to define a quality framework for sensor data.
* The implementation of a specialised monitoring system (quality indicators and paradata) and of soft and hard quality checks.
* The development of new strategies and methods for handling measurement (sensor) errors (i.e. missing data).

For sensor data, a quality framework is presented introducing different metrics - internal and objective metrics (intrinsic characteristics of sensor data) and context-based metrics - and two types of DQ estimation: (i) DQ assessment, which estimates the quality of the raw data; (ii) DQ evaluation, which estimates the quality of processed data considering context-based metrics.

For DQ assessment, many dimensions are considered, such as: believability, consistency (over time), timeliness (delay), accuracy (deviation from true value) and precision (granularity of readings). While for DQ evaluation, the need to acquire information on elementary units - paradata and contextual data - during the data collection phase and monitoring is highlighted.

The chapter also covered some topics - related to monitoring and processing building blocks - such as:

* Quality indicators, paradata and contextual data.
* New sources of measurement errors that can potentially influence sensor quality.
* The complex problem of missing in sensor data (with a short description of the strategy proposed by Bähr et al., 2020, for handling missing values in GPS sensor).

***Machine Learning***

Machine learning algorithms can be exploited in the statistical production process in performing the tasks of classification or regression of variables of interest. The functions that these algorithms can perform in a smart survey range from support to data collection phase to the data processing.

In particular, ML models can be applied to deal with input data provided by smart devices - such as images, signals, voice - during data collection, and for mixing/fusing sensor data, for the treatment of errors (anomalies in the data, missing, etc) in the data processing phase.

In the chapter, the use of ML is analyzed with reference to the architectural framework and building blocks: “Smart data tools”, “Smart survey monitoring” and “Smart data processing”.

The part on smart data tools discusses how a Machine learning model can reside directly on the devices where they receive the user's data (in-app) or be distributed as services and reside in centralized servers (in-house). The choice between the two solutions depends on the advantages and disadvantages provided by each approach and the requirements of the survey.

The smart data monitoring focuses on ML algorithms useful for exploiting the data produced directly by the device and to check data quality. In particular, this section discusses about:

* Quality checks that can be done with unsupervised ML anomaly-detection algorithms on paradata.
* ML tools for data quality monitoring through ML systems that learn from human feedback (Human-in-the-Loop, Active Learning, etc.).

The smart data processing blocks highlight different use of ML tools: (i) for the transformation of raw sensor data (signals, images, text) into statistical information (data preparation and data modelling); (ii) for the treatment of different types of errors in sensor data; (ii) for the assessment and evaluation of data quality.

**4.3. Privacy Preservation**

For the enhanced framework with respect to privacy, the ideas and concepts developed in the 3.1 and 3.2 were aligned with the work presented in the architecture sub-task. Most notably, first, the relationship between privacy preserving measures and GSPM phases was described, which can be summarized as follows:

* Identify Needs: privacy requirements should be determined based on the sensitivity of the data.
* Design & Build: implementations of privacy preserving measures that help to meet the identified requirements are developed along with the survey. Importantly, the development of these measures should be done in conjunction with the development of the survey as the implementation of these measures has a big influence on the way the survey is designed.
* Collect: execution of the chosen implementation starts. Without exception, this is before respondent data leaves their respective domain. Measures could range from pseudonymization to secret sharing.
* Process & Analysis: applied privacy preserving measures make sure that data is used in a way that follows the privacy requirements. For example, a data scientist looking at pseudonymised data will not be able to deduct any private information. Similarly, no employee or system at an NSI holding shares of respondent secrets will be able to extract private information during processing according to MPC protocols.
* Disseminate: the same considerations apply as do to all published statistics.

Second, for the architectural building blocks described in 3.3.1, the potential privacy issues that might arise were described, as well as measures required to deal with these.

* Smart Data Tools: on the respondents’ device, apps and external sensors can be used without restriction. When the data leaves the device, privacy considerations apply and requirements need to be met.
* Smart Survey Monitoring: these building blocks can be difficult to implement in a privacy-preserving manner, as they require respondent data at a highly granular level. Information content in these data is high and so has the potential to be highly sensitive. Advanced privacy preserving measures combined with clever design is required to deal with these when privacy requirements are strict.
* Smart Data Acquisition: active acquisition through questionnaire-like data collection should not lead to problems, as it does not differ from traditional questionnaire based surveys. Passive acquisition through sensors provides data with higher information content than provided by traditional means. As such, storage and processing of these data in accordance with privacy requirements might require more sophisticated measures.
* Smart Data Processing: efficient implementation of processing building blocks in accordance with privacy requirements requires strategic timing of processing steps. E.g. on-device pre-processing of a time-series of GPS coordinates into journeys defined by simply a start and end location, such that input data to aggregate statistical analysis is in a convenient format.

**4.4. Smart metadata**

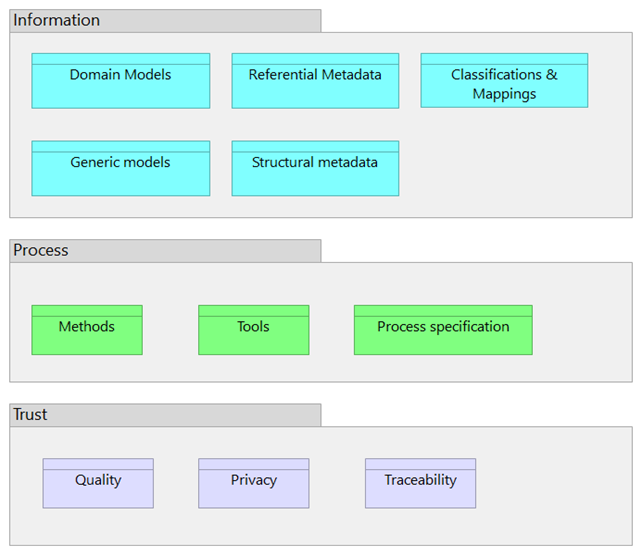
The business layer of the TSSu platform modelled the metadata management as an overarching business function, supporting and documenting the whole statistical process.

In order to investigate such a vast subject, in the preliminary framework, the metadata component was designed to meet the following general requirements:

* Compliance and alignment with official statistical standards and existing frameworks.
* Capture of metadata to describe the statistical process, including two dimensions: process standardization and reuse of application components:
* Use-case-driven approach, to test the initial assumptions related to metadata concepts selected for the preliminary analysis of the TSSU framework.
* The main component of the building block “Smart metadata management”, is the central Metadata Repository (MR) shown in the following figure and composed of three main areas: Information, describing survey main concepts variables, units, data structures, and classifications involved in data processing.
* Process, concerning all the tasks executed to produce a statistical output the methods and tools used for data processing, as well as a description of the main process steps to perform (Process specification).
* Trust, including several types of metadata, such as quality indicators and additional information to manage and monitor privacy issues, data provenance and process traceability. The elements grouped in this area are overarching, i.e. related to each step of the statistical process.

The metadata produced during data processing by a building block can be stored locally, or made accessible to the other building blocks through the central “Smart metadata management”, i.e. through the Metadata Repository. Thus, to run a procedure each building block may retrieve available metadata centrally stored in the MR.

*Figure 2: Metadata Repository subsets*



**4.5. Technical requirements**

The goal of the technical infrastructure is to show the possible ways of building the smart survey platform, including the technical representation of components used in different aspects of the smart surveys. It includes machine learning, metadata and privacy preserving technologies. The core part of the technical infrastructure is the technical requirements for components. General requirements show the possible ways of providing the minimal viable product.

More detailed requirements contain information with respect to the smart data tools – in device sensors and smart data acquisition – data storage and data transmission. For data acquisition different scenarios are presented – i.e., unstructured or structured repositories, according to the type of the smart data (machine generated vs. text) or the way the data is transmitted, e.g., async or sync – bulk or direct. The technical infrastructure shows also the four key components of the smart surveys platform, i.e. data storage, data access components, datalab and network security components.

Relevant information from the technical infrastructure part is also the characteristic of possible sensors used for data retrieval grouped by different devices, i.e. wearable, smartphones or tablets. It shows also the method that can be used to process the data.

Main findings from the task related to technical infrastructure show that unified platform for all smart surveys is possible, however it must be modular based on different ways of API to provide the data and storages (structured and unstructured) to store the information. It could be difficult to have one data storage, e.g. relational, to gather all the information from each smart survey as some of the smart surveys may deliver processed data, some of them raw data (e.g. pictures of the shop bills). To sum up – the key message is that the Smart Surveys Hub will allow to maintain unified technical components for each smart surveys, with respect to its specification – whether the data is processed in-device or at platform or the data provided is unstructured or structured.

Developing a smart survey platform is not a task on building the central repository but also a framework for all the applications sharing this repository. This is the reason why we included in this subchapter information on general requirements and possible data sources presented here as a set of possible sensors.

**4.6. Enhanced framework at work**

The section dedicated to the section Enhanced framework was conceived to analyse the deployment of the functionalities delivered by the building blocks described in the previous chapter. In order to identify pros and cons of different operational modes, the following user stories provided an overview of the interaction between the TSSu platform and the NSI’s environment.

*A - Active data acquisition through in-device sensors*

In the first use case, a National Statistical Institute accesses the TSSu platform to run a smart survey and collect data through a mobile app. According to the data collection strategy, the type of sensors and the smart data source, each NSI may download from the platform a set of software components. For example, the NSI could download and install in-house the following software solutions:

* Smart data collection tools, e.g., mobile app frontend, survey questionnaire design (variables, classifications, hard/soft checks, questions sequence and flow, etc.)
* Mobile app backend, e.g., a set of rest APIs that allow to store app data in a central repository and manage active data acquisition
* Monitoring system, e.g., a web application that offers a set of reports based on the analysis of the data stored in the Mobile app backend. The monitoring dashboards could, for example, monitor respondents’ participation and quality of collected data.

*B - ML training models through Input privacy preserving techniques*

In the second user story, assuming that smart data are stored in-house, or even during data gathering, NSIs can use the services offered by the TSSu platform to process collected data and train a ML model without sharing data. The TSSu platform may provide shared services and a distributed environment to execute ML training models by applying Input privacy preserving techniques, without sharing national data.

# Final remarks: main findings and key messages

The definition of the architectural and methodological framework for the design of a generalized European platform for Trusted Smart Surveys was a very challenging and stimulating task because the issues were completely new to official statistics and raised numerous questions, some of them still open. The first evidence from WP3 outputs is the very high complexity of designing a generalized and domain agnostic platform for smart surveys.

The most problematic aspects are probably those connected, on the one hand, to the protection of privacy, on the other hand to the management, treatment and evaluation of the quality of large amounts of data collected mainly through sensors.

Below, some final remarks and the main takeaways deriving from the activity of WP3 sub-tasks are reported.

**5.1 Methodological aspects**

Methodology concerns smart surveys with respect to many points of view, from data collection to data processing using machine learning algorithms, to the use of paradata and contextual data for evaluating and improving data quality.

In defining a general data collection framework for the TSSu, several dimensions that can affect respondent participation and data quality must be considered. Minimize the risk and burden on participants while maximizing the quantity and quality of data is of primary importance. The set of the sensors used can play an important role in the outcome of a survey, as data quality is intrinsically constrained by the characteristics of the sensors and the interactions of the participants with those sensors.

***Data collection strategies***

WP2 used the pilots to empirically address hypothetical problems during data collection. By doing so, in-depth insights and conclusions regarding respondent engagement, involvement of interviewers, and required survey monitoring could be drawn. The latter was also instrumental to understand what paradata is needed to make sense of the data collected or collection strategies used.

Data collection aspects, like monitoring or the use of contextual data (paradata), have not been practically experimented in a generalized perspective beyond the pilots. However, the lessons learned did impact the modelling of the TSSu building blocks and what should be considered during the data collection phase as best practise. In particular, when to have a human in the loop for assistance or when domain knowledge is crucial to assist data processing and ML procedures. In general, the interaction with the data providers (respondents or third parties) is a crucial focus, together with the different types of sensor data and providers, which has implications for privacy preserving and data processing (ML procedures), such as:

* bias caused by an incentive strategy, because respondent’s behaviour can be influenced by feedback and in particular incentive-based feedback mechanisms.
* new sources of representation and measurement errors, which imply the development of new data collection strategies aiming to prevent and control possible sources of error and of new methodologies in the assessment and correction of different types of error.

***Data quality***

A general data quality (DQ) framework should be declined right from the data source, the type of data and sensor, but also considering new sources of representation (coverage, participant selectivity, non-willingness to provide sensor data) and measurement (sensor) errors. From the platform perspective, the components involving methodology have been described and the implications discussed.

The main suggestions are the following.

* The application of the total survey error framework can provide a useful tool to guide methodological and practical decisions in smart surveys. However, it needs to be redefined, taking into account the hybrid forms of data collection and the device features for collecting, linking or processing data (device intelligence, internal and external sensors, data donation).
* A quality framework (DQ) for sensor data should be utilized, taking into account characteristics and properties of sensors and the quality of measurements.
* A monitoring system, based on quality indicator and paradata, should be defined for controlling the data collection phase and possible sources of errors. This building block can be generalised, but it requires specific functions for the different types of data and sensors.
* Functions or procedures are required to collect contextual data on the app usage and on performance of sensors, useful for assessing and evaluating the quality of sensor data. This part of data collection implies caution regarding privacy and minimization issues.
* Appropriate methodologies for handling with new types of data and new types of errors are needed.
* The strategies of dealing with missing items are very complex because data vary across sensors, depending on the extent and nature of the missingness patterns, and the phenomena under study.
* Smart surveys are generally mixed-mode surveys, using also “non-smart” modes, so issues regarding mode-effect should be taken into account, in order to guarantee or assess the representativeness of the estimate. Additionally, when respondent is actively involved in the data collection, some forms of bias, such as the social desirability in a traditional survey with interviewers, have to be taken into account.

***Machine learning***

Machine learning algorithms play an important role in smart surveys, as new solutions for the collection phase and for processing data should be explored and exploited, much more than in the traditional surveys.

The main key messages on this topic are the following.

* A generalized ML component has been developed, following an agnostic approach, to process the signals from different types of sensors and infer several variables, for example the physical activity or the means of transport of the respondent:
  + It was showed then that it is possible to develop a generalized smart data tool to transform signals into statistical variables
  + A ML component is generalizable in the sense that it can perform its functions in several contexts and in different phases of the survey.
  + The trade-off between the level of generalization and the quality of model is a crucial issue of this component.
* In-app or in-house solutions depend on the advantages and disadvantages provided by the two approaches and the requirements of the survey. The impact of each solution on training and re-training activities, privacy preservation (i.e into in-app solutions, user’s raw data does not leave the device and only the required information that is essential for the investigation is processed and centrally stored), and interaction with the respondents must be considered.
* In-house ML models do not necessarily have to be lightweight, due to the use of server’s resources, such as memory, or the scalability on cloud architectures. This can have positive effects on inference, accuracy and quality.
* ML applications can require human feedback in different functions to improve model accuracy and reduce data errors.
* The machine learning should be widen in terms also of online learning and active learning, including the respondent involvement.
* Concerning the (re)training of machine learning algorithms, further investigations are needed, due to the different context in each country. How often and at what stage of the survey a (re)training is needed are open issues.
* During the data collection, it is necessary to monitor the collected data. In the cases in which the quality of the data is not satisfactory, it is possible to refer to two kind of approaches: active approach (e.g. notification mechanisms to the user that require his action) or passive approach (e.g. through a centralized edit and imputation phase).

**5.2 Technical infrastructure**

Main findings from the task related to technical infrastructure show that:

* Smart Surveys Platform may be based on a unified set of components, including integration of various data storage and data acquisition types.
* There are lots of potential technical components which were not explored yet (e.g., specific sensors) but can be used by smart surveys for data retrieval.
* Data transmission in smart surveys platform should consist of various forms, i.e. async, bulk or direct transmission, according to smart surveys specification and requirements; however, it should be integrated by the use of different modules.
* Data storage should not be limited to one model; it should allow to use different modules, such as unstructured or structured, according to the requirements of smart surveys.
* General requirements for a Smart Surveys Platform should be stable and were formulated as a part of technical infrastructure.

**5.3 Architectural framework**

Comparing the initial assumptions with the achieved outcomes, the architectural task has shown the feasibility of process standardization and integration between traditional and smart survey pipelines.

This analysis is essential to specify more in detail the technical requirements of the TSSu platform, and start the development activities.

In addition to the final results, the main lessons learned can be summarized as follows:

* The pipeline for collecting and processing smart data sources needs to be customized according to the type of sensors used for data gathering
* Development of interoperable services through the adoption of official statistical standards, such as GSIM to harmonize data and metadata models
* The environment required for applying Input privacy techniques has a relevant impact on the design of the TSSu platform and further exploration is needed to assess the pros and cons and identify the best technical solution
* In order to integrate smart data pipeline and traditional survey tasks, some aspects, such as logistics and data collection monitoring, need to be adjusted according to the type of data sources

**5.4 Preservation of privacy**

Due to the nature of smart data, some smart surveys might require additional privacy preserving measures. In other cases, traditional measures might suffice. The following points are the main key messages with respect to privacy preserving:

* State-of-the-art technologies exist that allow for statistical analysis of respondent data in a completely oblivious manner. These technologies can mitigate the privacy risks involved in the use of highly sensitive respondent data. Implementation of these technologies in the context of actual surveys is, however, very complex.
* Application of advanced privacy preserving measures, such as secret sharing based MPC and processing in secure enclaves, interferes with the design of surveys and the way they are conducted. Therefore, privacy cannot be an afterthought, as the measures need to be developed in tandem with the rest of the survey.
* A general platform architecture that enables the application of required privacy preserving measures contains only a few additional components when compared to a platform without any of these measures. Most importantly, it requires a secure computing environment (secure enclave or distributed MPC servers) and a form of distributed storage. Such components require that the TSSu platform is (in part) distributed over involved parties.
* For the implementation of oblivious analysis, open-source software exists. However, implementation of these techniques in a production setting is a challenge and requires a team of experts in the fields of cryptography, statistics and software engineering.
* Application of privacy preserving measures impact all GSBPM phases, as such, development of these measures should go hand in hand with the development of surveys.
* Similarly, privacy preserving measures are not part of a single architectural building block for the platform. Rather, implementation of these measures should extend over all building blocks.
* Scalability of advanced privacy preserving techniques can be a concern. For example, certain non-parallelizable operations can take unfeasibly long in a MPC scenario; implementations of federated learning raise concerns regarding the amount of data transferred. As such, it is recommended that techniques are chosen and implemented with a given scale in mind, as scaling up is often non-trivial.

**5.5 Incentive schemes**

The task on incentives explored how gamification of surveys impacts survey design. Its main assumption was that incentives should be an integral part of the survey itself. Thus, gamification requires a redesign of a survey that fits state of the art design for app assisted surveys and offers a familiar experience due to its gamified nature.

The key messages resulting from this task are as follows:

* Gamified surveys convert hard tasks into smaller achievable goals, which is potentially beneficial for longer or longer running surveys.
* Gamified surveys rely on mechanics that offer an immersion into the activity, which is based on a strong integration into the user interface and challenges tied to actions that the respondent is asked to do.
* Gamified surveys must consist of different phases and offer new challenges over time to allow for continuous immersion.
* Gamified surveys must offer relevant feedback based on respondent’s actions.
* Gamified surveys should be designed with in-depth knowledge or at least abstract assumptions of the target audience.
* Gamified surveys could be easily combined with other incentives via a game-bases economy. For example, by offering a monetary incentive based on the number of completed challenges.

**5.6 Metadata management**

The metadata task has confirmed the feasibility of a central repository for the collection and management of the process metadata produced throughout data acquisition and processing. The metadata subsets, within the central repository, allow to monitor a process step during its execution, providing a quick assessment of the results and fostering the adoption of a metadata driven approach.

The main lessons learned concern the following aspects:

* The development of metadata-driven application components requires a detailed analysis of the functionalities provided by each solution. Even if the metadata task has confirmed the feasibility of process metadata standardization, unexpected peculiarities of some process steps may highlight the need of further adjustments of the conceptual framework. As an example, a particular data source may require ad-hoc data processing during data preparation, thus limiting process standardization.
* In order to implement shareable and trusted smart solutions, metadata management should allow to track also data transformations, Input/Output data structures and document the methods applied in each step.
* Metadata management should include paradata and analysis of contextual data captured during data acquisition, to provide auxiliary information for the quality assessment of collected data.

Main general conclusions

The overall findings of WP3 activities concern the issue of *generalization versus specificity*: Eurostat proposed the idea a platform independent from the domain and WP3 outcome is that this is not feasible on the whole, something can be developed in a generalized way but not everything. One proposal could be starting from developing some single modular solutions for some building blocks of the proposed architecture that can be considered shareable. Moreover, modularity of the solutions is the central point to balance generalization and specificity, thus fostering the implementation of reusable components.

Designing an omni-comprehensive platform for realizing smart surveys, not domain dependent, is therefore extremely complex and not needed for all aspects involved. However, responding to this complexity by developing domain-dependent applications, closely related to surveys, may not be a winning strategy in the long run. In a few years, we would have developed stove-pipe applications non-interacting each other, even if they could have some common elements. Further exploration is needed to identify the best strategy for dealing with this trade-off. This approach pushes the idea that we should develop the applications for the smart surveys flexibly, clearly distinguishing between what is (or should be) general and what is strictly domain-dependent.

Definitely one of the most important result of the WP3 is the description of a general framework for smart surveys, addressing the involved survey processes from different perspectives (methodological, technological, architectural, etc.), providing a new and useful reference scheme for the NSIs.

Some general conclusions and open issues regarding the main addressed topics are the following.

* From the architectural perspective, a solution regarding the integration between the management of sensor data coming from a passive data collection and the general survey process has been outlined but not designed yet. In the follow-up projects the two parts should be integrated and an end-to-end solution could be developed.
* From a technical point of view, although a set of potential software solutions have been suggested, the organizational challenge is still open. For ML, for example, how to use the training set, how to jointly manage the procedure in the NSI has to be organized.
* A central issue of defining a platform, in whatever sense, for smart surveys, is the open source matter: shareable solutions should be developed following the principle of open sources.
* Technology is not the main issue: the requirements of the technological infrastructure can be specified only when other issues are solved (e.g. privacy and respondent engagement).
* One of the main challenges of smart surveys is collecting passive sensor data from personal devices in a privacy protecting way. This is the blocking step as the two key rules of a smart survey concern respondent consent and data minimization. Moreover, the generalized application of privacy techniques is very challenging, as the right techniques have to be applied in the right case, requiring the right specific skills.
* From a methodological perspective, although many issues have been deepened through the interaction with WP2, some questions remain open and to be experimented:
  + Regarding smart data acquisition, the active-passive trade-off and the engagement and involvement of the respondents.
  + Regarding smart data processing, in-device versus in-house processing, with respect to privacy implications and data quality.
  + The trade-off among quality, burden and privacy preserving of data collected by a smart survey.
* Finally, the metadata task plays a crucial role in order to have a proper sharing of solutions, as metadata provide information about data sources, implemented tools, methods and description of data processing steps. Considering metadata management, important steps have been taken through the design of a central repository of process metadata, which should be enhanced according to surveys experts.

From the general discussion with WP2 and Eurostat, it emerged that the development of the Smart surveys deserves and requires further research and fieldwork, mainly for the issue of the involvement of the respondent, who must trust the NSI when sharing passive sensor data without having control over what he shares. NSIs have yet to test the feasibility at national level especially from the engagement and privacy/legal points of view. To this end, the availability of standard shared solutions could help NSIs experimenting these crucial aspects of the smart survey process.

1. Deliverable 3.1 Report on the Preliminary Framework, available from: https://ec.europa.eu/eurostat/cros/content/wp-3-conceptual-framework-european-platform\_en [↑](#footnote-ref-1)