Claudia De Vitiis

Trusted Smart Surveys: European Platform’s Methodological Framework

Claudia De Vitiis,[1], Mauro Bruno[1]; Jacek Maslankowski[2], Nils Meise[3], Joeri van Etten[4];

[1] Italian National Statistical Institute, Italy
[2] Statistics Poland, Poland
[3] Federal Statistical Office, Germany
[4] Statistics Netherlands, The Netherlans

Abstract:
The development of Smart Surveys offers new challenges to improve the quality of social surveys in the NSIs. The ESSNet on Smart Surveys, which started its activities at the beginning of 2020, will deliver preparatory work to create a European wide platform to share und re-use smart survey solutions and components. In this context, the work-package 3 is working on the idea of TSS platform through the conceptualization and development of a methodological and architectural framework for trusted smart surveys, following a top-down design approach. This paper focuses on methodological and architectural challenges related to smart data sources, together with privacy preserving features.

Keywords:
Architectural framework, Privacy preserving, Sensor data, Smart surveys, Trusted Smart Surveys

1. Introduction: The Trusted Smart Surveys and the ESSNet Smart Surveys

The Smart Surveys are surveys in which respondents are asked to employ smart devices (e.g. smartphones, tablets, activity trackers) to collect survey data through active and passive data collection, while in Trusted Smart Surveys (TSS) the respondents are also asked to share existing data collected by trusted third parties, like government authorities and larger, stable enterprises willing to establish data delivery agreements. These innovative way of data collection offer new challenges to improve the quality of social surveys in the NSIs. Constituent elements of a trusted smart survey are the strong protection of personal data based on privacy-preserving computation solutions, the full transparency and auditability of processing algorithms.

The ESSNet on Smart Surveys, which started its activities at the beginning of 2020, constitutes a contribution towards many important achievements foreseen within ESS: (i) testing and developing (trusted) smart surveys within the ESSNet, (ii) the conceptualization, development and implementation of a new reference architecture for trusted smart statistics as well as the evolvement of new skills within the ESS. The ESSNet will deliver preparatory work to create a European wide platform to share und re-use smart survey solutions and components. The platform will be agnostic to particular application domains, flexible and modular, implementing a set of common functions and configurable services that can be used to build particular instances of trusted smart surveys for specific application domains and/or target areas.

In this context, the work-package 3 (WP3) is working on the idea of TSS platform through two main tasks: (i) conceptualization and development of a general platform for trusted smart surveys, following a top-down design approach; (ii) development of proofs-of-concept in the form of modular prototype elements for essential aspects of the architecture.

In the following, the main features of the activity the work-package are described, in order to provide an overview of the different aspects of the framework, focusing in particular on the methodological (section 2.1), architectural (section 2.2) and privacy preserving aspects (section 2.3).

The framework will be compliant with existing frameworks such as GSBPM, GSIM, CSPA, SPRA and EARF. The target architecture will model smart statistical processes resulting from the combination of traditional and new data sources (e.g. sensor data, geolocation). As part of the definition of the TSS platform, a specific line of work is dedicated to the specification of the metadata that are useful in the context of smart surveys and special attention will be given to data structures and flows, process execution and traceability, and privacy, together with the representation of sensor data.

2. THE AREAS OF THE CONCEPTUAL FRAMEWORK

2.1. Smart Survey Methodology

The development of Trusted Smart Surveys based on mobile devices offers new opportunities for social surveys to collect data, thus generating new and unfamiliar types of data that are not standardized in structure, format, or availability, and requiring a sound methodological basis. Due to the widespread use of mobile devices, the growing number of available sensors in the device itself and via connected devices, they are an obvious choice for smart survey data [1, 2]. Usually these devices have one or multiple sensors that collect data for their own functionality or provide access to their sensor data to other applications. In addition, wearable devices and IoT devices for home automation became commonly available for consumers, which allows for even wider range of data access.

However, a lot of smart data sources are not created with the intent to be used in social surveys and do lack information needed by statisticians or social scientists. The base case for smart surveys is hybrid forms of data collection, where sensors assist, validate or replace survey questions. This also balances the tradeoff of smart surveys, which sacrifice exact questions to often ambiguous sensor readings, but make surveys more robust due to cross validation, offer ease of use and mean less burden for the respondents.

Machine Learning (ML) algorithms are essential tools when dealing with this new unfamiliar type of (sensor) data and its amounts. These innovative ways of handling sensor data raise issues related to the quality of data, such as training models and metrics to evaluate ML algorithms.

The collection of sensor data poses multiple challenges: selectivity of the participants, (non-) willingness to provide data, privacy and ethical issues, quality and usefulness of data, etc.. All these aspects have consequences both in terms of representation and measurement errors.

Representation errors can be caused by coverage issues. For example, ICT skills and digital literacy are not evenly distributed in a society or a sample population and provide a burden for the access to and acceptance of smart surveys. In addition, willingness to consent to passive data collection varies per type of sensor and depends on the context and purpose of the measurements. The more intrusive a sensor measurement is, the more respondents will refuse and the larger the potential damage of missing sensor data will be. Hence, introducing serious representation errors. Whereas measurement errors are mainly due to the sensors involved in the data collection process. Sensors measure a physical quantity and convert it into a signal, which a human or machine can process. Processed signals are categorized and meaning is applied to conceived patterns.

The evaluation of errors and their treatment are an important part in the methodological framework for TSS. Traditional/ legacy surveys try to cover this with the Total Survey Error, which can be expressed by the sum of row error, column error and cell error. When dealing with sensor/ big data the traditional error frameworks are limited, because non-standardized data management and transformation processes obfuscate parts of the process. Hence, the best approach to access the process is to evaluate the quality of the end product.

The methodological framework does not limit its scope on the survey, it includes respondents and their needs as well. In particular their need not to participate in voluntary surveys that are dull, offer no particular benefit and are easy to quit. Within the conceptualization of the framework incentives play an important role. Incentives shift the burden from the respondent to the survey agency, where they have to be implemented and taken care off. The challenge is to engage respondents and keep them engaged during the survey period. Common incentives (e.g. money, access to survey results) often fall short here, because they either offer an incentive before or after the successful completion of a survey. A powerful incentive to keep respondents engaged during in activity is to gamify it. To gamify a survey means to apply game mechanics to motivate and engage respondents in the survey [3, 4, 5]. Many mobile apps already include forms of gamification, which means respondents will have a familiar experience.

New data sources, use of ML, new errors, and new ways to create incentives for surveys are the aspects the methodological framework aims to cover by delivering guidelines, infrastructure and staff profile requirements for smart surveys, hybrid data forms and survey automation already in the design process.

2.2. Architecture and Technical infrastructure

The design of a European platform for Trusted Smart Surveys is a complex task that involves the analysis of architectural, technological and security aspects. In the first part of the ESSnet project we analyzed the business layer of the platform, using BREAL as a reference architecture [6]. BREAL divides business processes of the architecture into different phases, i.e. acquisition and recording, data wrangling, data representation, modelling and interpretation with shape output as the last phase.

The second phase of the project is focused both on the definition of the technical requirements of the platform and on the design of the platform’s information and application components, regardless of a specific domain or a particular survey. To achieve this goal, we analyze different architectural scenarios, combining the components involved in typical TSS processes. The following table shows the components that we combine in the architectural scenarios.

Scenario input	Description
Type of data acquisition	Passive (Centralized services for passive data acquisition)/ Active (Interoperable service for App implementation) Traditional/Smart data sources
Type of data provider	Respondent /Third party
Data Processing	In-app/NSI/ Platform/Third party

Data storage	In-app/NSI/ Platform/Third party
Service deployment	Interoperable/Shared/Replicated

Table 1: Architectural scenarios inputs

The scenarios allow to identify strengths, weaknesses and opportunities related to the different workflows, more precisely our analysis focuses on:

- Data processing: one of the key capabilities of the TSS platform is related to data processing. Depending on the scenario and / or on national regulations, data could be processed on mobile devices (edge computing), in the NSI infrastructure or directly in the European platform. In the latter case challenging IT problems should be faced, i.e. privacy-preserving methods and advanced cryptographic techniques (secure multi-party computation)
- Data flow: another key aspect concerns data storage and data flow from respondent devices to NSI infrastructure or the European platform. Depending on device's sensor, data can flow can be continuous (NSI or European platform data processing), asynchronous (device processing) or bulk (data is pre-processed on device and sent to NSI or European platform).
- Service deployment: services can be deployed both in NSI infrastructure and the European platform. Depending on the use cases, NSIs may choose to run the services offered by the platform locally (replicated services), e.g., advanced machine learning algorithms implemented at European level, running in NSI infrastructure.

As an example of the scenario analysis performed in WP3, **Errore. L'origine riferimento non è stata trovata.** shows the Archimate [7] model of the scenario "*Passive data acquired through shared or replicated services in the NSI's infrastructure and processed in the European platform*". In this scenario the NSI uses a smart data service offered by the TSS platform. More precisely the NSI installs within its infrastructure the smart data service to collect smart data provided by respondent's devices. Collected active data is sent to the TSS platform where it is processed by smart data processing services.

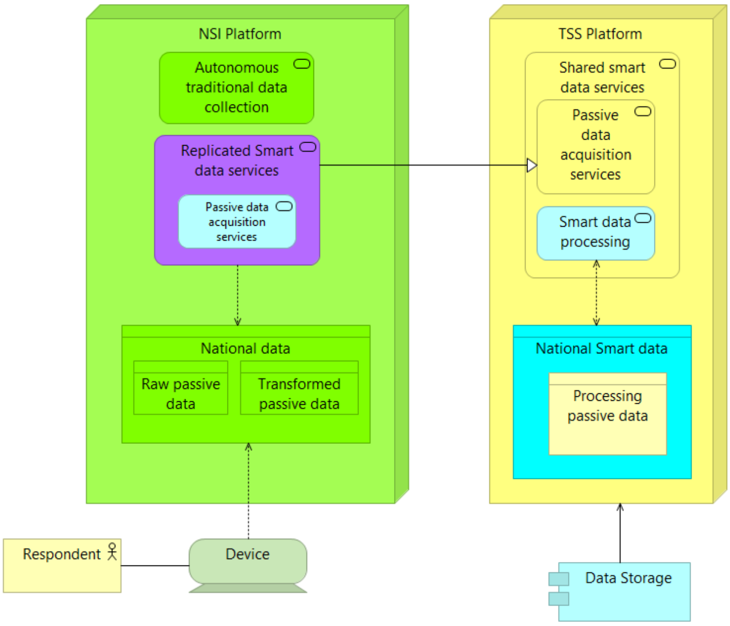

Figure 1: Passive data acquired through shared or replicated services

Later we will focus on the technological aspects of the components involved in the architecture scenarios, e.g., data storage (structured, unstructured files vs. SQL or NoSQL databases) or data transmission channel (synchronous, asynchronous or event driven).

2.3. Privacy preserving

To maintain full respondent's trust and be fully compliant with the EU and national privacy regulation ESS Smart surveys need to apply established principles, develop guidelines and introduce new technologies and techniques in area of privacy and transparency. This task explores best practice, identifies concrete challenges related to Smart surveys, and will proposes guidelines as well as methodological and technical solutions.

Guidelines will be based on established privacy-by-design principles such as seven principles listed below [4]: 1. Proactive not reactive; preventive not remedial; 2. Privacy as the default setting; 3. Privacy embedded into design; 4.Full functionality – positive-sum, not zero-sum; 5. End-to-end security – full lifecycle protection; 6.Visibility and transparency – keep it open; 7. Respect for user privacy – keep it user-centric.

Processing the Smart surveys input personal data will also require new architectural, methodological and technological techniques and approaches, for example: Architecture that doesn't need to centralize sensitive data at a single entity to reduce concentration risk (for example Edge computing); Advanced privacy-preserving methods for particularly sensitive data such as anonymization and advanced cryptographic techniques (for example Secure Multi-Party Computation); Solutions for full auditability and complete transparency of the processing methods applied to the data.

An important consideration when the aforementioned techniques are applied in an official statistical setting like the ESSnet Smart Surveys, is whether these interfere with the statistical process. For example, if data is pre-processed considerably for minimization purposes, so-called contextual data is lost, which could have helped in explaining anomalous entries. Hence, minimization of data for privacy protection could, in this example, interfere with the statistical process. Also, when cryptographic techniques such as secret sharing or homomorphic encryption are used, restrictions on e.g. the possible mathematical operations and allowed datatypes are introduced. These restrictions could limit the usability of the smart survey data, hence, interfering once more with the production of statistics. Given these considerations, a focal point of the task will be to investigate these trade-offs and find a configuration that strikes the optimal balance between privacy and quality statistics.

Combined, these principles, guidelines, techniques and related considerations will assure high level of trust in private, decentralized and automated collection as well as processing of (sensor)data, for usage in ESSnet smart surveys, while simultaneously retaining the level of quality required of official statistics.

3. Discussion and Future work:

Some aspects of the framework are currently under development in Proofs-of-Concept, with the aim of testing modular prototype elements for essential aspects of the architecture. These experimentations focus in particular on: (i) machine learning and trade-off between active and passive data collection, (ii) incentive schemes implementation, (iii) metadata modelling, (iv) privacy preservation and (v) architectural and IT components of the platform. For some of these aspects the proof-of-concept consists of identification and implementation of use cases.

References:

1. T. D. Buskirk and C. Andres, “Smart Surveys for Smart Phones: Exploring Various Approaches for Conducting Online Mobile Surveys via Smartphones,” *Survey Practice*, vol. 5, no. 1, 2012.
2. F. Keusch, B. Struminskaya, C. Antoun et al., “Willingness to Participate in Passive Mobile Data Collection,” *Public Opinion Quarterly*, vol. 83, pp. 210–235, 2019.
3. J. Harms, S. Biegler, C. Wimmer et al., “Gamification of Online Surveys: Design Process, Case Study, and Evaluation,” in *Human-Computer Interaction - INTERACT 2015,* J. Abascal, S. Barbosa, M. Fetter et al., Eds., pp. 219–236, Springer International Publishing, Cham, 2015.
4. K. W. R. Oliveira and M. M. V. Paula, “Gamification of Online Surveys: A Systematic Mapping,” *IEEE Transactions on Games*, p. 1, 2020.
5. T. Triantoro, R. Gopal, R. Benbunan-Fich et al., “Personality and games: Enhancing online surveys through gamification,” *Information Technology and Management*, vol. 121, no. 2, 2020.
6. ESSnet Big Data II, Work package F, deliverable F1 “BREAL: Big Data REference Architecture and Layers Business Layer”. [On-line] Available from: https://ec.europa.eu/eurostat/cros/sites/default/files/WPF_Deliverable_F1_BREAL_Big_Data_REference_Architecture_and_Layers_v.03012020.pdf
7. ArchiMate R 3.0.1 Specification, Open Group Standard [On-line] Available from: https://pubs.opengroup.org/architecture/archimate3-doc/toc.html.