|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | 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**

[Link to our CROS website](https://ec.europa.eu/eurostat/cros/content/essnet-smart-surveys_en)

**Workpackage 2**

**Smart Survey Pilots**

**Deliverable 2.5: Shareability of smart surveys in the ESS**

**Version 1.2, 24-05-2022**

**Prepared by:**

Annemieke Luiten (CBS, Netherlands)

Patrick Lusyne (StatBEL, Belgium)

Barry Schouten (CBS, Netherlands)

ESSnet co-ordinator:

Workpackage Leader:

Barry Schouten (CBS, Netherlands)

[Jg.schouten@cbs.nl](mailto:Jg.schouten@cbs.nl)

telephone : +31 70 3374905

SUMMARY: WP2 performs four diverse pilots to inform WP3 on the specifications of a smart survey platform in the European Statistical System (ESS). Important questions are to what extent design and architecture of smart surveys can be shared across ESS countries and what are country deviations. The answers provide insight into the shareability and required flexibility of smart survey platforms.

Following the description of smart survey pilots, as given in deliverable 2.7, a distinction is made between CSPA, methodology, logistics and ethical-legal levels.

To answer the main questions, extensive functional testing has been conducted by NSI’s participating in WP2.1 Consumption and WP2.2 Time use. For WP2.3 Health and WP2.4 Living conditions testing has been exploratory and does not cover all design elements of a smart survey. Furthermore, the number of countries in these two sub-WP’s is smaller. As a consequence, some building blocks and design levels are not discussed or only relatively general and abstract. Even so, first conclusions are drawn.

INTRODUCTION

In deliverable 2.7 it has been explained and elaborated that a smart survey design consists of various elements that can be chosen differently across NSI’s:

* IT frontend:
  + Questionnaires and diaries
  + Linkage to internal and external sensors
  + Data donation
  + Data processing
* IT backend:
  + Web server
  + Data models
  + User database
  + Case management
  + Linkage to existing data
  + Data processing
* Methodology:
  + Recruitment material
  + Data collection strategy
  + Plausibility checks and validation rules
  + Sensor data machine learning models
* Logistics:
  + Interviewer recruitment/motivation of n
  + Monitoring and analysis
  + Helpdesk (email, phone)
  + Landing page website

The elements are termed building blocks in the application layer described in WP3 deliverables. They are a mix of applications, human interventions and other supporting functions. Their implementation also determines the type of smart survey scenario as discussed in deliverable 3.3 of WP3.

A certain minimum number of the building blocks/elements need to be common in order to assure a viable shared solution in an ESS context. Viable means that 1) resulting statistics are comparable in time across the ESS, and 2) tools can be maintained and updated efficiently across countries. Shareable smart surveys, thus, concern at least interoperable or replicated services.

ESSnet Smart Surveys aims to determine if there is sufficient common ground and how this would translate to an overarching architecture in terms of IT, methodology and logistics. It tries to do both top-down in WP3 and bottom-up in WP2. WP2 performs four smart survey case studies and in the design of these surveys collects the requirements of the participating countries. From the country inputs, generic and country-specific requirements are extracted.

This deliverable reports the findings on four levels: CSPA, methodology, logistics and legal-ethical. These are briefly described in the next section. Next, the generic and country-specific requirements are pooled into frontend and backend features. Finally, consequences for WP3 are discussed. It is imperative to distinguish these two components. The backend infrastructure may be largely national, although being harmonized in terms of metadata and data models, while frontend tools may be shared and maintained centrally.

LEVELS IN SHAREABILITY

In elaborating and documenting smart surveys, four levels were identified:

1. CSPA conceptual, logical and physical levels
2. Methodology level
3. Logistics level
4. Legal-ethical-policy level.

CSPA stands for Common Statistical Production Architecture and represents conceptual, logical and physical levels of smart survey tools. Eurostat maintains an inventory of tools for the Household Budget Survey (HBS) and the Harmonized European Time Use Survey (HETUS) at

<https://webgate.ec.europa.eu/fpfis/wikis/display/ISTLCS/INVENTORY>

The methodology level concerns methodological design decisions:

* User interface of frontend: The most influential design choices in the user interface.
* Data collection strategy: This concerns the use of contact modes, contact and reminder strategies, incentive strategies, recruitment materials, the use of “non-smart” modes.
* Data quality checks: Soft and hard checks of plausibility of entered data and notifications of missing data.
* Machine learning models in processing or mixing/fusing sensor data

It is customary that NSI’s make their own choices in some of the methodological decisions. The relevant information is whether they affect implementation of the smart survey.

The logistics level refers to parts of the input stage that require human interaction or intervention:

* Recruitment: Many of the ESS surveys have an interviewers-assisted data collection such as doing the starting questionnaire, recruiting and assisting respondents in a diary, motivating respondents during data collection, and/or picking up closing questionnaires.
* Monitoring: Sample units may not participate, drop-out or deliver insufficient data quality. Monitoring dashboards may be inspected on a frequent basis in order to determine whether interventions are needed at overall level.
* Assistance: Respondents may be assisted in starting and using smart survey tools passively through a helpdesk and website or actively through interviewers and or technical experts.
* Human-in the-loop machine learning: In sensor data applications, models seldom reach 100% accuracy. Certain population subgroups or certain survey statistics may require manual inspection. In ESSnet Smart Surveys where feedback of statistics to respondents is deemed important, such human-in-the-loop processes may even occur during data collection.

Like the methodology level, countries can make their own choices, but the implications for frontend and backend need to be derived.

The legal-ethical-policy level refers to what is legally allowed, ethically accepted and policy-wise decided. The legal level in essence refers to Data Protection Impact Assessments (DPIA) and describes minimal legal requirements. In the ESS, the GDPR (General Data Protection Regulation) is the starting point for the legal level, but countries may have additional legislation. The ethical and policy levels essentially represent country-specific requirements that are imposed. In the ESSnet, the focus is on smart surveys and it is sufficient to look at what is new, relative to existing survey data collections:

* The data collection may use the computing and storage options of personal devices
* Part of the data collection may be passive
* Part of the data collection processing may be shifted to the personal devices rather than in-house at the institute
* Part of the collected data may be of specialist content of which part of the respondents has no knowledge, even under informed consent
* In order to be able to participate, one may have to possess a specific device or to accept that one needs to use a device that is provided

The ESSnet established a working group legal that divided smart surveys into in-device data storage, in-device data processing, processing by other parties than the NSI, and in-house processing. The working group produces a set of recommendations and will most likely also make an application at the EDPB (European Data Protection Board). In this deliverable, only first conclusions are drawn.

RESULTS OF FUNCTIONAL TESTS

Country requirements are summarized to:

1. Generic: These concern features and functionality that all NSI’s share. They are considered to be MUST haves;
2. Country-specific majority: These concern features and functionality that the majority of NSI’s prefer and that do not impact strongly comparability. They are considered to be SHOULD haves;
3. Country-specific minority: These concern all other features and functionality that are preferred by a few NSI’s and do not affect comparability. They are considered COULD haves;

In order to be shareable, a smart survey application must satisfy generic requirements and preferably acknowledge the requirements specified by the majority of ESS countries. On the contrary, requirements expressed by a minority of countries may threaten shareability. The consequences in terms of IT may be that a tool must be unrealistically flexible and very intensive to maintain and keep up to date. In terms of methodology they may be that UI’s become incomprehensible, machine learning models too complex or inaccurate, and surveys too burdensome to respondents.

It must be noted that in the specifications data donation of bank transactions data and other personal data and the use of points-of-interest databases such as streetmaps are not included. These additional smart features are, however, explored and may be added at a later stage in follow-up projects and/or joint activities.

*FRONTEND SHAREABILITY*

Generic

* CSPA:
  + All: All platforms and operating systems of at least five years back are supported
  + All: Multi-device display and access is possible; desktops/laptops only if appropriate
  + All: Multiple household members can participate
  + All: UI can consist of one or more questionnaires and a diary, and questionnaire answers can be used in the diary
  + All: Apps can be downloaded from country app stores with country-specific labels and descriptions
  + Consumption: Minimal output on amounts of all purchases within the COICOP classification according to ESS regulations
  + Consumption: Option to both scan shopping receipts and enter data manually by respondents. Scans that are detected as not being receipts can be refused and the respondent alerted
  + Health: Accelerometers must be the same brand and type in the sample
  + Living conditions: Indoor climate systems must be the same brand and type in the sample.
  + Living conditions: Respondent questionnaires are needed to provide context on dwelling/life style through a questionnaire
* Methodology:
  + All: Tutorials and instructions on data entry are made accessible through the UI and clearly describe the anticipated active-passive engagement of respondents
  + All: The UI should conform to common practices in application design and is responsive
  + All: Individual feedback/insights are provided. If desirable from a measurement perspective, feedback is provided at the end of data collection, rather than during.
  + Consumption: Use is made of detailed and common language country product lists to support manual data entry
  + Consumption: Product search strategy is configurable on a number of features (distance function between two strings, lower threshold in string distance, maximum number of results displayed, historic frequency sold in shops). Respondents can add new products that are not available in the list
  + Consumption: In-app feedback is given on the quality of scans of receipts and respondents can decide to redo scans
  + Health: Accelerometers must be research-grade and worn on the same part of the body. Sensor technology must be documented and open.
  + Living conditions: Indoor climate systems must be objectively reported/evaluated to measure accurately with known sensor errors and known calibration and machine learning procedures. Sensor technology must be documented and open.
* Logistical:
  + All: Helpdesk and support options are presented and directly linked to inbound call centers and technical helpdesks
  + All: In case interviewers are involved in recruiting and motivating respondents, then they should be well acquainted with the app and be able to answer the most common FAQ about the app UI
  + Health: Clear instruction material is necessary to guarantee adequate and proper use of accelerometers
  + Living conditions: Clear instruction material is necessary to guarantee adequate and proper implementation of climate systems including a technical helpdesk/support option
* Legal-ethical:
  + All: Information on the usage of respondent data following GDPR is clearly explained either in the UI or on the linked landing page
  + All: All platforms are supported and operating systems to at least five years back
  + All: Respondents should be able to submit data without using mobile device sensors or camera
  + All: Respondents must be able to check and edit data during or after the fieldwork period
  + All: In case interviewers are involved in recruitment, they should be able to explain how data are stored and processed
  + Consumption: Respondents can cut scanned images in such a way that no confidential data are submitted
  + Health: Respondents can refrain from wearing the sensor for periods of time they want to keep private.
  + Health: Respondent data have to be transmitted directly to the NSI backend.
  + Living conditions: Respondent data have to be transmitted directly to the NSI backend.

Country-specific majority

* CSPA:
  + All: Country NSI logo and colour scheme, country language, country currency, default user interface language (e.g. EN or FR) for other languages than country language can be tailored
  + All: All questionnaire modules are accessible within one user interface
  + All: Paradata on navigation behavior can be logged
  + Consumption: Additional details on number/metrics of all purchases, country of purchases
  + Consumption: Intro questionnaire has country-specific content
  + Consumption: Option to include digital receipts
  + Consumption: Plausibility checks can be included
  + Consumption: Info on quantities/metrics of products can be provided
  + Consumption: In app cropping of images can be turned on/off
* Methodology:
  + All: Use and display of incentives in the UI (if applicable)
  + Consumption: In-device processing of receipts and respondent validation is possible and respondents can edit results
  + Consumption: An in-app tutorial movie is added in order to explain how to scan a receipt and select the relevant part of the receipt
  + Consumption: The type of store or store name can be extracted or entered by the respondent in order to aid receipt language processing
  + Consumption: In-app image processing can detect low quality scans and advice respondents to retake a picture
* Logistical:
  + All: Country-specific helpdesk/support per NSI linked to the user interface
* Legal-ethical:
  + All: Option to login-logout from the application at any given time

Country-specific minority:

* CSPA:
  + All: Multi-language options within one country are possible
  + Consumption: Additional details on yes/no biological/ecological purchases, type of shop/shop name of purchase, household member who made purchase, yes/no online purchase
* Methodology:
  + Consumption: Option to apply a filter to select purchases for specific household members and/or specific dates
* Logistical:
  + Consumption: Option to contact the interviewer assigned to the household through the UI
* Legal-ethical:
  + All: Option to anonymize submitted data within a household
  + All: Option to alter passwords for the application

*BACKEND SHAREABILITY*

Generic

* CSPA:
  + All: Device-switch is possible and synchronization available across multiple devices
  + All: The backend is linked to a monitoring/analysis server
  + All: The backend coordinates one or more questionnaires and a diary
  + All: During the fieldwork period respondents can edit/revise submitted data
  + All: Apps can be used off-line
  + Consumption: Automated processing of receipts is installed, i.e. the backend is linked to a receipt processing server
  + Consumption: The backend can detect submission of scans that are not receipts and return a warning to the respondent
  + Health: Data classification is linked as a service to the backend.
  + Health: Accelerometer raw data should be available. This leads to research-grade devices such as ActivPal, ActiGraph, UKK
  + Living conditions: Data fusion and classification is linked as a service to the backend.
  + Living conditions: Indoor climate system raw data should be available
* Methodology:
  + All: General nonresponse metrics of registration and completion relative to background characteristics are available
  + Consumption: Alternative modes for data entry are supported and connected to the case management system
  + Consumption: Different machine learning options are available depending on the type of available data (ranging from no data, i.e. full annotation, to linked printed receipt texts, i.e. string matching)
  + Consumption: Online learning for classification of scanned receipts is implemented
  + Consumption: Performance metrics of receipt processing are provided and can be evaluated
  + Health: Generic (machine learning) algorithms are available to classify intensity and type of activity
  + Health: Generic (machine learning) algorithms are available to classify the main indoor air parameters.
* Logistical:
  + All: Data collection coordinators need to be able to monitor progress and status of responses, preferably by linking to a specified number of background variables
  + All: Helpdesk need to be able to consult status of individual cases
  + All: Case management allows for switch of modes, e.g. paper or interviewer-assistance
  + All: In case interviewers are involved, they can be informed about progress and status at any given time
  + Consumption: Machine learning models for receipt processing can be evaluated by a text mining expert i.e. active learning
  + Consumption: Machine learning model training for receipt processing options (ranging from no data, i.e. full annotation, to linked printed receipt texts, i.e. string matching) are documented and available to text mining experts
  + Health: Clear and detailed protocols are needed for handling, downloading and transmitting respondent data to the backend, including data completeness and plausibility checks.
  + Health: Machine learning routines are updated, and if necessary retrained, through active learning at least every three years
  + Living conditions: Machine learning routines are updated, and if necessary retrained, through active learning at least every three years
* Legal-ethical:
  + All: Only the information needed to derive household statistics is maintained
  + All: Apps can be used off-line and synchronized at any given time
  + Consumption: Only the information needed to derive household expenditures is transmitted and stored, i.e. excess information printed on receipts is removed
  + Health: Only classified data are transmitted for further processing. Raw sensor data are isolated from other data.
  + Living conditions: Only classified data are transmitted for further processing. Raw sensor data are isolated from other data.

Country-specific majority

* CSPA:
  + All: Paradata on in-device navigation can be stored in the backend database
  + Consumption: country-specific performance metrics of in-device and/or in-house processing of receipts are stored in the backend database
  + Consumption: The backend URL is country-specific, i.e. data are stored and handled in local NSI web servers
* Methodology:
  + All: In-app paradata on navigation behavior can be evaluated through summaries that are informative of technical errors, lack of respondent motivation and/or respondent inability to interpret in-app tasks
  + Consumption: Performance of in-device and/or in-house processing of receipts can be evaluated by each participating NSI
  + Consumption: The in-house processing of receipts can detect receipts from countries other than the country of the NSI and present them for manual inspection
* Logistical:
  + All: Multiple contact modes can be applied, i.e. invitation letters, email, text messages
  + All: Statistical queries evaluating in-app paradata can be consulted at any time
  + Consumption: Manual check/intervention of processed receipts is possible during the fieldwork period
  + Consumption: Overviews of products on receipts unknown to machine learning/matching procedures can be listed and fed back to training of models
* Legal-ethical:

Country-specific minority:

* CSPA:
  + All: Altering of individual passwords is supported
* Methodology: -
* Logistical:
  + Consumption: Interviewers need to be able to login to the backend to consult the status of their households
* Legal-ethical:
  + All: Respondents can change their password

CONSEQUENCES FOR WP3 POC’S AND ENHANCED FRAMEWORK

This section explains implications for WP3 in terms of main focus areas an in important building blocks.

WP3 develops proof-of-concepts (PoC) for four modular prototype elements: architecture and metadata, machine learning models, privacy preservation and incentives/gamification. The PoC elements are deemed most complex and prominent in platform specifications and are explicit focus areas in WP3. To these four areas, the general trade-off of active-passive data collection is added.

TABLE 1a - d: Shareability for each of the POC elements plus active-passive trade-off per pilot

|  |  |
| --- | --- |
| Pilot | Architecture and metadata |
| Consumption | * All: A prominent role is played by receipt scanning, evaluation and processing * All: Unknown products can be logged and stored * Majority: In-app and in-house receipt processing metrics should be logged and stored * Majority: In-house processing of receipts can be automated |
| Health | * All: dedicated software needs to be made available. * All: metadata need to be logged and stored |
| Living conditions | * All: A dedicated cloud environment needs to be secured, which can be centralized or country specific. * All: dedicated software needs to be made available. * All: metadata need to be logged and stored * All: smart survey need to support questionnaires to provide context on dwelling/life style |

|  |  |
| --- | --- |
| Pilot | Machine learning |
| Consumption | * All: Receipt text extraction and classification methods are imperative * All: Receipt text extraction is largely based on pre-trained methods and can be implemented with relatively simple country dependencies * All: Receipt text language processing largely follows the same rules across countries but it is recommended to include shop-specific (and thus country-specific) rules for discounts and use of multiple lines per product * All: Receipt text classification is strongly country-dependent and various methods can be applied depending on the type of available data * Majority: In-app OCR and language processing is helpful in respondent interaction * Majority: A minimal form of online learning for receipt classification is imperative and country-dependent |
| Health | * All: data cleaning algorithms; it is imperative that data handling is the same for all countries * All: machine learning algorithms to determine MVPA; it is imperative that these are the same for all countries |
| Living conditions | * All: data cleaning algorithms * All: machine learning algorithms to determine IEQ and underlying causes * All: Respondent questionnaires are needed to provide context on dwelling/life style through a questionnaire |

|  |  |
| --- | --- |
| Pilot | Privacy preservation |
| Consumption | * All: Given that receipt product classification cannot take place in-app, at least the pre-processed receipt texts need to be submitted * Majority: Receipt scanning can be organized such that only products and prices are stored and submitted. However, it is recommendable that respondents can edit text extraction results * Majority: In-app OCR/language processing metrics are crucial to understanding scan quality |
| Health | * All: The measurement devices only register movement, no location. Measures are stored on the device. The device is sent back to the NSI. There is no sensitive information on the devices. |
| Living conditions | * All: choose a measurement device that stores information on the device or otherwise secure the data transmission to the dedicated private European or NSI cloud environment. |

|  |  |
| --- | --- |
| Pilot | Incentives and gamification |
| Consumption | * Majority: In-app insights into expenditures led to slightly higher recruitment rates and slightly lower drop-our rates. However, differences are not statistically significant * Majority: There is some evidence that monetary incentives increase recruitment rates |
| Health | * (unclear as yet how this generalizes):   + evidence in one country that the monetary incentive amount does not have an important impact on recruitment rates   + evidence in one country that promised feedback as incentive does not have impact on recruitment rate, and may even have a negative impact on some groups. * Feedback on physical activity patterns is appreciated by participants, but can only be given at the end of measurement |
| Living conditions | * All (?): the insights on IEQ are a major incentive for participants. * All: insights need to be given at the end of the measurement period only. * Recruitment rates (evidence in one country) may be high (depending on how the recruitment question is framed), even without offering a monetary incentive |

|  |  |
| --- | --- |
| Pilot | Active-passive trade-off |
| Consumption | * All: A minimal respondent interaction on receipt scan quality is imperative * Majority: In-app OCR and language processing can be helpful * Majority: In-app interactive insights on expenditures are helpful * Minority: Respondents can edit basic results of OCR ad language processing |
| Health | * All: in the device chosen for the pilots, no active role for the participants is possible; no data are transmitted before the end of the measurement period. Participants could be asked to fill in a diary to provide context to the measurements. |
| Living conditions | * All: in the device chosen for the pilots, no active role for the participants is possible; data transmission is only one way, from measurement device to cloud. Participants could be asked to fill in a diary to provide context to the measurements. |

WP3 distinguishes functions and actions in their GSBPM based business layer and building blocks in their application layer. Having shareability in mind, what functions and actions are most demanding and what building blocks are most crucial?

Demanding functions:

* Active/online learning: In virtually all smart surveys underlying classifications and raw data are subject to mutations over time;
* Local storage and processing: New operating systems demand for checks and revisions of mobile app code;
* Product search in HBS: Product/store lists to support product search string matching procedures: Lists should have sufficient diversity and richness to guarantee usability. For countries this is a new element;
* Receipt scan language processing in HBS: Languages show subtle differences in receipt format which requires small revisions of code when adding a new country;
* Receipt product classification in HBS: Product classification is strongly country dependent, though machine learning model training may largely be copied;

Most crucial building blocks:

* User interface for respondent interaction on sensor data errors/deficiencies;
* (Semi-)automated machine learning model re-training procedures;
* Infrastructure and logistics for external sensor systems;

DISCUSSION

Based on the considerations and summaries, tentative conclusions can be drawn about the necessity to ensure shareability about each of the levels. Shareability is a trade-off between comparability, efficiency and flexibility. Countries are inherently different in survey climate and survey case management, so that ensuring comparability demands for some flexibility. However, flexibility adds complexity to IT and methodology maintenance and updating. Also, there is a breaking point where comparability in statistics is at risk.

The following MOSCOW assessment is made of shareability of the four levels:

* MUST
  + Methodology level in both frontend and backend
  + Legal-ethical level frontend
* SHOULD
  + IT level frontend
  + Logistics level frontend
  + Legal-level backend
* COULD
  + IT level backend
  + Logistics level backend

Methodology and frontend legal-ethical specifications must be shared by countries. Methodology is the main input to guarantee equivalence of concepts and handling of errors in survey data, sensor data and other forms of data. Since legal-ethical constraints determine what the tool presents to respondents and what processing of data is done in-device versus in-house, also these should be shared by countries.

Preferably, the frontend IT and logistics and the backend legal-ethical constraints should be shared as well. The frontend IT determines how survey questions and measurements are implemented and how respondent interaction is organized, say to validate and adjust sensor measurements. There are multiple ways to design UI’s and country-specific user experience may be different. However, differences should be modest to ensure measurement equivalence. Logistics linked to the frontend such as helpdesk support and other supporting facilities should ideally be equally well elaborated across countries. The backend legal-ethical constraints mostly concern the handling of the different types of data. These may be different based on backend solutions, but should satisfy the same principles.

Finally, the IT backend and IT logistics are the most flexible and could be shared. It is here where the country differences come in most strong, while impact on comparability of statistics may be smallest.

In follow-up research and evaluations, these tentative conclusions on shareability should be revisited.