|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | 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 2.12: End report**

**Version 1.1, 28-06-2022**

**Prepared by:**

Annemieke Luiten, Barry Schouten (CBS, Netherlands)

Patrick Lusyne (StatBEL, Belgium)

ESSnet co-ordinator:

Workpackage Leader:

Barry Schouten (CBS, Netherlands)

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

telephone : +31 70 3374905

Table of contents

[1. INTRODUCTION 3](#_Toc107382311)

[2. GUIDE TO WP2 DELIVERABLES 4](#_Toc107382312)

[3. MAIN RESULTS AND LESSONS LEARNED 5](#_Toc107382313)

[3.1 Time use 7](#_Toc107382314)

[3.2 Health 10](#_Toc107382315)

[3.3 Living conditions 12](#_Toc107382316)

[4. SMART SURVEY LEGAL LEVEL 14](#_Toc107382317)

[4.1 PRESUMPTIONS IN LEGAL CONSIDERATIONS 15](#_Toc107382318)

[4.2 FIRST RECOMMENDATIONS 15](#_Toc107382319)

[4.3 THIRD PARTY INVOLVEMENT 16](#_Toc107382320)

[4.4 TYPE OF DATA PROCESSING 17](#_Toc107382321)

[4.5 OPEN DECISIONS IN SMART SURVEY DESIGN FROM THE LEGAL PERSPECTIVE 18](#_Toc107382322)

[4.6 NEXT STEPS FROM THE LEGAL PERSPECTIVE 20](#_Toc107382323)

[5. THE WP3 FRAMEWORK AT WORK 20](#_Toc107382324)

[5.1 WP3 GSBPM STAGES AND BUSINESS LAYER 20](#_Toc107382325)

[5.1.1 What GSBPM stages have been investigated? 20](#_Toc107382326)

[5.1.2 What GSBPM sub-stages are particularly complex and/or time-consuming? 23](#_Toc107382327)

[5.1.3 What GSBPM sub-stages have become mature? 24](#_Toc107382328)

[5.2 WP3 APPLICATION LAYER AND BUILDING BLOCKS 25](#_Toc107382329)

[5.2.1 What building blocks have been developed? 25](#_Toc107382330)

[5.2.2 To what extent can building blocks be re-used? 26](#_Toc107382331)

[5.3 CLASSIFICATION OF WP2 CASE STUDIES THROUGH WP3 SCENARIOS 27](#_Toc107382332)

[6. NEXT STEPS AND RECOMMENDATIONS 31](#_Toc107382333)

SUMMARY: This is the WP2 end report. It provides a guide to the deliverables, summarizes the main results and lessons learned and provides next steps and recommendations. It also includes the conclusions and considerations of a legal-ethical working group that was established during the ESSnet in order to bridge the gap between survey designers, mobile app development and legal officers. To go from lessons learned to next steps and recommendations, the WP3 smart survey framework, as described in deliverable 3.3, is applied in this end report to the WP2 case studies.

1. INTRODUCTION

Work package 2 is the empirical and applied component of ESSnet Smart Surveys. It has identified the needs, (further) designed and tested smart survey solutions in four main survey areas: Consumption, Time use, Health and Living Conditions. Consumption is explicitly linked to the Household Budget Survey (HBS). Time use is explicitly linked to the Harmonized European Time Use Survey (HETUS) but also to more specific time use themes. The focus in Health has been on physical activity tracking, which may be linked to the European Health Interview Survey (EHIS) and national counterparts that are conducted in between EHIS years. The focus in Living Conditions has been on indoor climate, which may be linked to Statistics on Income and Living Conditions (SILC), again to EHIS and also to housing surveys conducted in most countries. For themes Consumption and Time use it employed existing app-based solutions that have been and are developed in parallel ESTAT-funded innovation projects. In considering the specific survey themes and designing/elaborating smart survey solutions it functioned as the model for WP3.

The objectives set out for WP2 at the start of the ESSnet were:

* Identification and collection of the functional and technical characteristics of the solutions currently being developed for the HETUS and HBS, and if available of other survey based data collections;
* Establishment and execution of a solutions evaluation protocol for existing data collection applications. The protocol should be used with an adequately chosen fit-for purpose sample, to further improve the solutions;
* Developing pilot projects demonstrating the use of scalable secure private computing solutions (e.g. Secure Multiparty Computation) to process individual citizen data without concentrating personal data at a single entity (no risk concentration), in combination with appropriate technological solutions to provide full auditability (e.g. non-modifiable logging of transactions).

and had the following activities in mind:

* Effective tactics to engage and involve “citizens”, i.e. persons and households in the ESS;
* Shareability of tools across ESS countries, i.e. identification of generic and country-specific features
* Modularity of tools for similar topics within themes Consumption and Time use
* Accuracy and comparability of (trusted) smart survey data relative to regular data collection
* Specifications for data collection and data processing infrastructure for WP3

These activities indeed have been part of WP2. However, investigation into effective tactics to engage and involve respondents has been limited to the case study on consumption/HBS. Consequently, also evaluation of accuracy and comparability have been limited, although also in te case study on physical activity tracking comparisons have been made to traditional survey approaches. The main reasons were discussions on legal boundaries and COVID-19 measures.

The prominent role of discussions linked to legal boundaries and requirements set into motion an unplanned working group on legal issues. In the working group both legal officers and survey designers were active.

An important anticipated task/activity of WP2 has also been the provision of specifications to WP3. WP2 formed the empirical and practical counterpart of the more conceptual and generic WP3. It was only when WP2 started to see results from the various tests that it could provide useful information to WP3. Especially in the last eight months of the ESSnet there has been frequent communication between the work packages.

This end report is structured along the following sections: In section 2, the WP2 deliverables are enumerated and briefly described. In section 3, the main results and lessons learned are summarized from the WP2 deliverables. Next, in section 4, the legal considerations are discussed. Section 5 puts the WP3 framework at work. Finally, section 6 ends with next steps and recommendations.

1. GUIDE TO WP2 DELIVERABLES

WP2 has three types of deliverables: 1) documentation and specifications, 2) test evaluations and 3) overarching views on modularity and shareability.

The documentation and specifications of WP2 applications were made based on a template comprising of four levels: IT architecture (following CSPA), methodology, logistics and legal-ethical. This template is given by deliverable 2.7. Based on the specifications, a mapping is included to the six smart features:

* 1. In-device/local storage and processing;
  2. Application of internal mobile device sensors;
  3. Application and linkage to external sensor systems;
  4. Linkage to online public databases;
  5. Data donation through the respondent;
  6. Data donation through the statistical institute;

Furthermore, a mapping was made to the main interest areas, or proof-of-concepts (PoCs), of WP3. The WP2 deliverables directly based on this template are:

* D2.9: Description of the HBS app following D2.7;
* D2.10: Description of the MOTUS application;
* D2.8a – d: Specifications to WP3, which are in part updates of D2.9 and D2.10 based on functional tests and field tests;
* D2.11: Product, shop and activity lists used in the HBS and MOTUS applications;

The deliverables on test evaluations concern conclusions drawn in usability tests and field tests. They concern the plausibility of substantive statistics and recommendations for the main survey design choices. The design choices again refer back to the four D2.7 levels. The deliverables directly based on tests are:

* D2.1b and d: Evaluations of the design choices varied in usability tests for time use and indoor climate;
* D2.1c: Evaluations of the design choices varied in usability tests and small-scale field tests for physical activity;
* D2.1a: Evaluations of the plausibility of substantive results for HBS from a large-scale field test in three countries (ES, LU and NL);
* D2.3: Evaluation of potential user interface design choices to improve respondent engagement in the HBS;
* D2.4: Evaluation of the role of interviewers in improving respondent engagement in the HBS;

The last set of deliverables takes a more general stand and looks at two crucial features of smart surveys in the context of the ESS: modularity and shareability. Modularity concerns the potential of smart survey solutions to be used in varying settings with closely related information needs. It is the basis for efficiency, given that smart surveys demand for considerable investments in IT and logistics. Shareability is about the application of smart survey solutions to the wider context of the ESS. It considers the common denominator in specifications for smart survey solutions. The deliverables falling under this broad viewpoint are:

* D2.2 a and b: Modularity of, respectively, the HBS and MOTUS applications;
* D2.5: Shareability of applications in the four WP2 case studies within the ESS;

WP2 deliverables can be found at

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

1. MAIN RESULTS AND LESSONS LEARNED

3.1 Consumption

Within the ESSnet, functional tests, usability tests and field tests have been linked to the Household Budget Survey app developed by NL. It is the only case study of the four that also conducted a large-scale cross-national field test to evaluate push-to-app strategies. The main results and lessons learned will abstract as much as possible from the specific implementation. Nonetheless, results and conclusions cannot be seen separate from the following ambitions and vision of an app-assisted HBS embedded in the design:

* The app should facilitate two smart features: advanced product search for manual data entry and automated processing for receipt scan data entry;
* Respondents should be engaged through personalized feedback on their expenditures;
* Respondents should be involved by allowing for editing of receipt scan processing results;
* The HBS has an in-app start questionnaire;
* The app is configurable in UI language, logo, start questionnaire, product and store lists, in-app OCR and language processing, and product search algorithm parameters, but not in the general look and feel and UI layout;
* Recruitment for the HBS can be multi-mode;

In order to design the UI and the smart features, both functional tests and usability tests have been conducted (see deliverable 2.1a and deliverable 2.5) during 2020 and 2021.

To evaluate the utility and format of personalized feedback and in-app editing of receipt processing results, experimental conditions were added to the 2021 field test (see deliverable 2.3)

To find empirical evidence for a multi-mode push-to-app strategy, the interviewer condition was added to the 2021 field test (see deliverable 2.4).

The following conclusions were drawn:

* App-assisted HBS:
  + In terms of frontend usability, an app-assisted smart HBS is very well feasible. During field tests there were but few technical issues and respondents looking for help;
  + In terms of backend, and especially back-office, an app-assisted HBS is demanding as it needs to be linked to existing case management and fieldwork procedures. It implies an investment either through a replicated service or an interoperable service, depending on the available knowledge and resources at an NSI;
  + Machine learning routines play a crucial role in receipt processing and must be viewed as a micro-service that requires separate maintenance and coordination;
* Smart features:
  + Product search algorithms are effective but depend on the richness and form of language of product lists. Creation of such lists facilitates also non-app implementations and requires an investment in time;
  + The receipt scan data entry option is added value; roughly a quarter of reported purchases came through scans. The quality of product extraction depended only mildly on length of receipts;
* Respondent engagement:
  + There is no strong sign of higher response rates, lower drop-out rate and/or improved data quality when personalized feedback is offered. This may to some extent be the consequence of how it was operationalized in the HBS app, but it confirms other previous studies. Nonetheless, in usability tests it was appreciated and it should be kept;
  + Interviewers turned out to take the role of an intermediary role when it comes to app use questions. It is recommended that they have used the app themselves for a few weeks;
* Respondent involvement:
  + The involvement of respondents in both product search and receipt scans has been one of the main focal points during development. It is advisable to perform usability tests;

Automated In-app OCR and NLP of receipt scans are feasible but should be supplemented by the option of in-app respondent editing. Respondent assistance in product classification is not recommended;

* Push-to-app:
  + Interviewers have a clear positive impact on response rates and the extra response shows similar data quality and attrition rates. Furthermore, they do play an intermediary role when it comes to app use questions. There are only small and inconclusive signals that response is also more representative;
  + It has yet to be evaluated and tested whether alternative modes in sequence to an app have a beneficial impact at representation while maintaining measurement equivalence;

Detailed conclusions on shareability across the ESS can be found in deliverable 2.5, on substantive plausibility in deliverable 2.1a and on smart survey design in deliverable 2.8a.

Overall, the most demanding new features of a smart HBS are machine learning routines for product classification and the creation of rich product lists.

Recommended next steps:

* IT/architecture level:
  + Semi-automated receipt processing needs to be perfected and elaborated
  + Add uploading of digital receipts as data entry option
* Methodology level:
  + UI for uploading of digital receipts should be developed and included through a series of usability tests
  + Multi-device/mode data entry should be possible
  + Improvement of ML approaches for classification is paramount
* Logistics level:
  + Elaborate active and online learning options of ML models
  + Create ESS-wide FAQ and support pages
* Legal-ethics level:
  + Improve access to the survey through multi-device/mode options
  1. Time use

Two pilots were dedicated to the time use theme. Both were qualitative with small study populations (a few tens of persons) but featured elaborated scopes covering a wide variety of substantive and methodological matters. To some extent, not having a large-scale field test became leverage to increase the degree of experimentation in these pilots.

The MO:TUS data platform[[1]](#footnote-1), annex mobile application developed by the Vrije Universiteit Brussel[[2]](#footnote-2) and Hbits played a vital role in both pilots. In the first pilot, organized during the summer and autumn of 2020 while the pandemic was peaking, several questionnaires on various subthemes (core time use, labor time, mobility) were operationalized in MO:TUS and tested by several participating National Statistical Institutes. The methodological focus of this pilot was on functionality, on evaluating whether MO:TUS is functioning adequately, in other words.

In the second pilot, the thematic interest shifted entirely to (passenger)mobility, and methodologically the aim was to evaluate certain aspects of MO: TUS' usability, how well MOTUS functions. This pilot had a much more elaborated scope than the first one as several surveys were combined with a diary and some smarter elements like notifications and a geofence. Also different was that Statistics Belgium operationalized this research flow in the MO: TUS' backend almost entirely autonomous of the Free University Brussels. This hands-on approach, and the elaborated scope, resulted in a completely different experience compared to the first pilot.

The many insights gained from the experiences in both pilots fall into various categories. Firstly some findings relate to the functioning of MO:TUS in general. In particular from the second pilot, there is also a substantive conclusion for the field of (passenger)mobility. Additionally, there are broader observations about the methodology of organizing smart surveys. Finally, working on these pilots also led to more general reflections about the field of smart surveys and how to advance.

The two pilots were not exhaustive evaluations of MO:TUS covering all aspects of the application functioning. Nevertheless, these experiments offered at least some insight into some of the application's mean features. Overall the conclusions of the functional tests are very favorable for MO:TUS. Technically the application functions well, and its basic features are fully operational. Almost no technical flaws were experienced in the operationalization or during the test of essential research elements such as the surveys and diaries. The issues that were detected seemed occasional and limited in nature.

Although the basic features of MOTUS like organizing a survey are fully operational, smarter features still need to gain maturity before they can be deployed in genuine research. This will take additional developments of the team developing MO:TUS at the Free University Brussels. For example, retrieving in the backend the geographical coördinates resulting from the tracking necessary to enable the geofence was not possible yet.

The pilots had no direct substantive-thematic objectives. However, passenger mobility, the central theme of the second pilot, has a substantive significance in relation to the challenges ahead that the Green Deal brings to NSIs. Despite its small size and limited scope, the result of the second pilot indicates that there might be an alternative to the highly fragmented production of passenger mobility statistics in the European Statistical System. By teaming up and deploying data platforms like MO:TUS, it looks feasible to organize a harmonized cross-national survey capable of measuring the most elementary mobility indicators in a harmonized way.

MO:TUS also rates reasonably high on the usability dimension. The participants in the second pilot were asked to rate how usable MO:TUS is using a questionnaire covering several usability dimensions. Although these participants had been subjected to a complicated research flow, overall ratings were fairly good (see DL 2.1). A majority of the participants, all staff of Statistics Belgium we have to add, also expressed the firm conviction that this kind of technology certainly must have a place in the future production capacity of National Statistical Institutes.

In the same deliverable, the idea of a third dimension to the spectrum formed by functionality and usability is introduced. How usable is MOTUS from the point of view of the statisticians, or in other words, how deployable is the platform? The hands-on experience in the second pilot gave an excellent opportunity to evaluate this. The outcome is again favorable for MO:TUS. Especially MO: TUS' capacity to enable statisticians (in this case of Statistics Belgium) to work directly in the platform's backend, through a GUI-like application developed on top of the MO:TUS backend, enhances MOTUS's deployability highly.

Two methodological, closely related insights result from the pilots. The first is the importance of communication, and the second is that research flows need optimization, especially when they combine various modes of data collection.

Both observations seem obvious, but their full extent becomes apparent only by experiencing them first-hand.

By deliberately increasing the complexity of the research flow of the second pilot, the communication challenge to guide participants through the research also increased drastically. The communication in place in the pilot failed to rise to that challenge, not only because of several sequences of bad communication in specific parts of the research but also because of the general approach that was not apt to cope with the challenges that come with organizing research via mobile phone. The conclusion of this minor communication 'debacle' is the profound realization that this kind of research necessitates new communication strategies rather than an adaptive approach based on current communication practices.

Adding smart elements to a research design increases its complexity, almost by definition. Hence, the importance of communication, but throughout the second pilot, it became clear that the logic and the order of research elements also was critical. Matching the functioning of active and passive elements, in particular, can be a challenge. Therefore, we must develop and optimize pathways and scenarios to guide respondents through the elaborated research flows that characterize smart surveys.

Experiencing first-hand the issues related to the organization of a study featuring a complicated research design that integrates active and passive elements triggered a conceptual awareness of the importance of active measurement in developing smart surveys. Active measurement through surveys or diaries allows contextualization, while sensors can support active questioning to alleviate response burden. This perspective results in a more stepwise approach to how National Statistical Institutes can modernize and 'smarten' their production capacity. Rather than aiming to take a direct leap forward to mainly sensor-based research, progressively elaborating the current production with mobile active data collection, and gradually supporting this collection with certain smart elements, seems more efficacious.

In our opinion, such a gradual approach may even change the current mindset concerning smart surveys and innovation. A more explicit emphasis on mobile and active features of smart surveys makes it necessary to involve other stakeholders, such as production statisticians. As they are the persons within national statistical institutes that manage the current production, their involvement ensures a much more effective integration of smart surveys into NSI's current production capacity and a greater legitimacy of that integration.

Recommended next steps: The above insights came at a price, as the ESSnet on smart statistics was challenging in many aspects. They imply steps forward in a various way. This experience not only deepened our understanding of smart surveys from a technical and methodological viewpoint it also made us reconsider the strategy ahead. Hence, the steps forward below are both concrete and strategic and represent several strands.

* Like mentioned above, there is a need for further development of certain smart features of the MO:TUS application. The Essnet on Smart Statistics had a valuable contribution in identifying specific gaps (e.g. in relation to the geolocation service). The MO:TUS developer team at the Free University Brussels is very dedicated, and new developments are underway (geolocation service, receipt scanning). An additional dimension is that these new developments will need elaborate testing. Therefore, Statistics Belgium and Free University Brussel are considering deepening formally or informally their collaboration in this field.
* As communication is so critical, expanding our expertise in that domain will be essential. This will require investments in experimentation on a larger scale than what we had in this ESSnet, or at least in the Time Use theme. Communication is a broad field, and future communication strategies should strengthen both recruiting and process through flow.
* The integration of active and passive measurement is undeniably a vital issue, and advancing on this must be one of the priorities in the coming years. In our vision, the further development and integration of the geolocation application in combination with survey and diary research should remain one of the principal axes of development to achieve this.
* Success in all of the above will depend on the quality of the tests, pilots and experiments that precede actual implementation. As both human and financial resources are limited international collaboration may optimize cost/benefit ratios of national testing strategies. Therefore, in the relatively short term, steps should be taken to continue the cooperation that has developed within the ESSnet.
  1. Health

For the topic of ‘health’, the attention was focussed on measuring physical activity by means of an accelerometer. It is a well-known fact that measuring physical activity by means of a questionnaire leads to gross over-estimation of physical activity. Hence, sensor measurement would potentially lead to higher quality data, if a representative part of the population would agree to participate. The accelerometer used in the pilots is the thigh-worn activPAL. This is a commercially device and comes with hardware, software to download the data from the measurement device in various data formats, software to visualize the data in various ways, and algorithms to interpret the data in terms of length and intensity of physical activity. The device stores the data, that are downloaded once the device is sent back to the NSI. Small scale feasibility pilots were performed in three countries. Three more extensive quantitative pilots were performed in the Netherlands. Some attention was given to the feasibility of using participants’ own devices. The quantitative pilots were aimed at determining uptake in the general public, and studying push-to-app strategies. All quantitative pilots were web-only, with invitations sent by email. The invitation was for filling in a questionnaire. At the end of the questionnaire, the request was made for follow-up research with the accelerometer.

An important consideration with these pilots and results, is that it concerns results of one possible solution: the activPAL accelerometer. Deliverable 2.1 details why this solution was chosen. However, other solutions and accelerometers are possible, with other implications for third party involvement, logistics, usability, privacy and ethics.

The following conclusions were drawn:

* Response rates and push to app interventions

The most important conclusion was that it is feasible to measure physical activity with accelerometers, that respondents are willing to some extend to participate, but that additional steps need to be taken to increase uptake and adherence.

* + In all quantitative pilots, around 50% of the respondents to the questionnaires consented to participate in the follow-up study with accelerometers. Depending on the initial questionnaire response rate, this amounted to 13% to 49% of the sample.
  + Some dropout will occur in the following stage: not all people who consent to receiving one will actually start wearing it, and not all people who start wearing it will do it for the requested 7 days.
  + Around 50% of questionnaire respondents owned a private accelerator or app. Around 60% of those were willing to copy data from their own device into a questionnaire, while 35 à 40% was willing to donate and upload their own data. The quality of the private accelerometers varies however, and people tend to adapt or at least round their data when copying them out. The actual number of respondents who donate their data is studied in the last pilot, but the amount is not yet known at the time of writing. We did find that it is not so very simple for some kinds of personal devices to upload the data, and we had to restrict ourselves to one kind only in the pilot.
  + Participants and people who have an accelerometer or use an app to monitor their physical activity are more active than those who do not participate and /or do not have a private device or app.
  + In one pilot an experiment was performed with promising feedback on physical activity. This promise had a small *negative* influence on participation rate, that reached significance for women.
  + Incentive experiments were performed in the pilots and incentives were discussed in the usability pilots. Various monetary conditional incentives were offered, varying from 12,50 to €25. In a related pilot, not within the ESSnet, €40 was offered. Although, as expected, higher compliance was attained with the higher incentives, the differences were small. In one pilot, respondents were asked what incentive amount would induce them to participate. The amounts mentioned were far higher than what would be feasible in practice: all above €40. In the last and largest pilot in the ESSnet, we worked with a promised incentive of €12,50.
* Machine learning
  + For understanding physical activity, two elements are relevant: the activities performed, and the intensity with which they are performed. For all pilots, the algorithms developed by activPAL were used. However, if objective measurement with accelerometers is implemented, ‘official statistics’ algorithms need to be developed; we can obviously not be dependent on commercial algorithms that are on the one hand not transparent, and on the other hand can change without our knowledge. Some initial effort was undertaken to train our own algorithms, but they are as yet not good enough. Careful discussion with data users needs to determine the variables that an algorithm must be able to deliver. Perfect recognition of all possible physical activities is probably less important than recognizing in broad categories that a person is sedentary or active, and with what intensity.
  + There are already international research collaboration platforms that work on the development of machine learning algorithms for thigh worn accelerometers.
* Statistics

The pilots in this WP on health were targeted at studying feasibility. Only preliminary analysis was performed into the substantive contents of the data in terms of physical activity and the manner in which we should use these data for the statistics on physical activity that are published by NSI’s. The preliminary study showed, as expected, that the objective measurement showed that people are far less active than they mention in questionnaires. Like was discussed for the ML algorithms, discussion with data users needs to determine the kind of statistics that needs to be developed, including new measures that objective measurement allows.

Recommended next steps:

* All pilots were performed in a web-only setting. Obviously, interviewers will have a major role in securing a higher uptake and higher adherence. Most probably, interviewers will (also) be able to increase participation in groups that are not automatically enthusiastic about participating in research with accelerometers. This will likely diminish the bias towards more active participants that was evident in the ESSnet pilots. So, additional field tests that involve interviewers are necessary to be fully informed about the possibilities, the participation propensity of groups that are not or less represented in web surveys, and the bias that rests after interviewer involvement.
* Additionally, field tests in other countries will inform about generalizability in the European context of the mostly Dutch findings.
* Discuss with data users the kind of data they would like to receive from this new source, and
* Subsequently develop proprietary algorithms to generate the requested variables.

* 1. Living conditions

Living conditions are related to national health surveys, the European Health Interview Survey (EHIS) and to the Statistics on Income and Living Conditions survey (SILC). In this work package, living conditions are limited to indoor environment quality (IEQ). Air quality, and more specifically air pollution has been defined by the WHO as one of the biggest threats to human health, alongside climate change (WHO, 2021). Mostly, attention has been given to outdoor air quality, but indoor air quality likewise has been related to a wide number of diseases, like asthma, lung cancer, cardiovascular diseases, irritation of eyes and skin, obesity and diabetes, but also do decrease of productivity and learning ability. Objective measurements are quasi impossible to be provided by respondents by means of questionnaires. Respondents will not be aware of and/or unable to report on aspects of indoor environment that may impact their health. Examples are CO, CO2 and fine dust (particulate matter), but also noise. Such sensor data measurements have been explored already in European funded research on school and office indoor environment, but are still new to official statistics. However, NSI’s are in a unique position to combine the subjective measurements in surveys, knowledge about respondents’ illness and health and registry knowledge on buildings with the objective measurements that sensors can provide. This project was hence more about new data than about new methods of data collection.

Very few respondents will have devices that can measure these substances, which means that devices to measure indoor environment will need to be provided to respondents. That means that these measurements introduce a significant cost-quality trade-off discussion.

In a number of small scale feasibility pilots in three countries, we focused on user acceptance and usability. In one large scale pilot in the Netherlands, panel participants were presented with vignettes in which the request to participate in IEQ measurement in the home was varied along different dimensions. In addition, the quality of the measurement device and data has been given considerable attention. Deliverable 2.1 details the choice for the measurement system (the uHoo air), the design of the feasibility studies and the quantitative vignette study and results.

The following topics were addressed in the pilots, and the following conclusions were drawn:

* Will people accept an IEQ measurement system in their home? What is the perceived respondent burden?
  + Depending on the question format, up to 61% (in a ‘yes/no’ question), and up to 77%[[3]](#footnote-3) (in a scale version) of respondents in a web questionnaire were willing to allow a measurement system in their home, with no other incentive than the feedback they would receive on their IEQ. Willingness is related to respondent characteristics like age, education, and worries about air quality. In a situation where interviewers could be involved in recruitment, both consent and representativeness will most likely increase to very high levels.
  + Once the device is installed, no further burden is perceived. Participants appreciated learning about their IEQ and even appreciated being shocked by some of the findings, leading to sometimes lasting behaviour change. Burden may be introduced by unexplained high levels of substances and by warning notifications.
* Data quality
  + There are some concerns about the quality of the data that have been measured. We have been unable to receive satisfying information from the distributer on measurement precision and reliability, nor on the calibration algorithms that are used for the sensors. Even though one study (Baldelli, 2021) showed satisfying measurement quality for some of the sensors (the other ones were not studied yet), this lack of openness is disturbing. We have found on the other hand, that the data analysis shows interpretable results with at least face validity. The consensus among the experts is that (cheap) sensors can give an impression of the relative variation over time. The question is, of course, if that is good enough for our purposes. Precise definition of those purposes will help answering this question.
* Logistics: are people able to install the device and the app? How much support is needed? Do they adhere to instructions?
  + for most participants the download of the app, the subsequent installation of the uHoo and connecting it to WIFI was quickly and easily done. However, a substantial minority of participants struggled and had to try several times, either by resetting the device or by restarting the app. This is a level of devotion that is probably not to be expected in the general public. An alternative approach to the installation, or the creation of an instruction video, or finding a device where the installation is easier needs to be embraced for future large scale measurements
  + Participant compliance to instruction is very important. Nevertheless, even the very dedicated participants did not always follow the instruction on the placement of the device, or the instruction to keep it connected to both electricity and wifi all the time.
  + Deliverable 2.1 gives in addition attention to the topics of where to measure, how long to measure, and the costs of measurement.
* Privacy issues, both in terms of objective criteria (who has access to the data, how do the data reach the NSI’s back end) and subjectively. The objective aspect of privacy is described in detail in deliverable 2.8. The subjective aspect was discussed in focus groups and questionnaires and is described in deliverable 2.1.
  + There is a difference between countries in the perceived sensitivity of the data, with about half of the German and Dutch participants judging the data as sensitive, against only 10% of the Belgians.
  + The most relevant element in the evaluation of the sensitivity of the data, according to focus groups in two countries, is the amount of context data that is linked to the IEQ data. Trust that nothing ‘hidden’ is measured is also a relevant dimension.
  + Data should ideally be stored offline on the device and not be sent via WiFi to a cloud, and third-party suppliers should be avoided or at least chosen according to their publicly perceived trustworthiness.
* Respondent behavior
  + It was found that about half of the respondents changed their behavior as a result of the feedback: mostly they aired more often and longer. Even though the respondent feedback proved to be an important element of the respondent enjoyment, it is advisable to give feedback only at the end of the measurement period, and not during. Behavior change does not need to be a measurement problem, if the behavior is enduring. The extent to which this is the case is however unknown at this moment.

Recommended next steps:

* One or more quantitative pilots with employment of interviewers to gauge the ultimate response propensity and logistical burden on the organisation.
* If possible, combine with measurement of outdoor EQ, and measurement in the work environment of the participant, in order to measure a participant’s total exposure.
* Investigate further the business case for these new data. Who are the (potential) clients? What is the balance of costs and gain?

1. SMART SURVEY LEGAL LEVEL

Legal restrictions and the GDPR context had a strong impact on the early phases of the ESSnet evaluations within WP2. This was especially true for WP2.2, Time use, where a third party was involved in design and data collection. During the ESSnet it was decided to install a working group on legal and ethical constraints and restrictions consisting of a mix of survey designers and legal officers. The working group produced an extensive report on what is new in smart surveys relative to traditional surveys and what open technical questions need to be answered. The report included a draft of a data protection impact assessment (DPIA) for smart surveys. The working group also produced a discussion document on two main design decisions in smart survey: third party involvement and in-device versus in-house processing. Here, this discussion is included as it is paramount that guidelines and best practices are found for future projects and initiatives. The working group remains active after the ESSnet and one main recommendation that may be moderated by the working group is a joint/simultaneous application to national legal authorities and/or the overarching European counterpart (EDPS).

* 1. PRESUMPTIONS IN LEGAL CONSIDERATIONS

Four presumptions were made by the working group as regards to the functioning of smart Surveys:

1. Smart surveys collect data that are entirely oriented at a specified and existent information need. Given that information need is much more abstract than the answers to questions in a corresponding non-smart survey, this must be interpreted as that they serve fully and only the concepts of interest in the information need.
2. IT security of data transfer and data storage, specifically that on respondent devices, follows common best practices and is implemented according to generally accepted norms confirmed by external security auditors. It must be stressed that NSI servers and data warehouses are much safer than are respondent devices.
3. NSI’s will not forward data/information to the respondent devices. This would pose a threat to these data, having the security difference between NSI and device in mind.
4. The respondent gives explicit consent to sensor data and/or data donation and can see the outcomes of measurements.

In other words, the data that are being collected through sensors or data donation are completely needed, there is an acceptable risk of a breach, they are not enriched with data of other persons/households and respondents have given consent.

* 1. FIRST RECOMMENDATIONS

Starting from the presumptions in section 4.2, the working group made first, tentative recommendations:

1. Taking into account that there are different National Statistical Institutes and Eurostat involved in the project, and the purpose of the processing is to elaborate European Statistics as set up by law, it is highly recommendable that all NSIs participants and Eurostat become Joint Controllers inside the project to better clarify the tasks and responsibilities among NSIs (also authorisation for transmission among all partners). It will also give more legal certainty to the users and citizens and increase the transparency and trust.
2. As long as personal data are collected for the purpose of a statistical operation covered by the European Statistical Programme or the National Statistical Plan, the legal basis for collecting is the EU and national statistical laws and, generally, article 6.1 e) GDPR:

* EU and National Law will apply to the respondents: It is obligatory for the respondent to cooperate and answer the survey or voluntary.
* For the processing of sensitive personal data (Article 9 GDPR) or for data that are not required by a statistical operation in the statistical programming instruments the consent of the respondents is always needed.
* The age of consent of minors should be the common foreseen in article 8.1 GDPR ( 16 years old).
* On the other hand, there is the consent of the interviewees for the app to access the different functionalities of their device (camera or location). In these cases, consent relates to secondary aspects, i.e. aspects in which, if consent is not given, the information can be provided by other means. For example, if a citizen does not want to give consent to activate the camera, they will not be able to take a photo of the purchase receipt and will have to manually enter the products they bought into the app.: In order to access the different functionalities, the consent of the interviewee must be requested.

1. Transparency and clarity are needed: To this end it is recommended:

* To send a letter to the household explaining everything necessary and providing the household's username and password to access the survey and the country to which their household belongs.
* When downloading the app and selects the country the respondent have to enter the user name and password and enter the session of his/her household, where all the members of the household would appear.
* Each member of the household would enter their user and the first time they enter they are asked to enter a password in order to protect sensitive data within the household
  1. THIRD PARTY INVOLVEMENT

The extent to which third parties involvement is regulated is one of the two main new dimensions, because access to personal data is limited to those that have a legal mandate for data collection for statistical purposes.

Four options for smart surveys as regards third party involvement are recognized:

1. There is no third party involvement
2. Third party involvement is both regulated and tailored to the specific need: It means that the third party is a processor and intervenes “on behalf” of the NSI that signed the contract and stated the specific need, including process of data collection, including security.
3. Third party involvement is regulated but not tailored: In this case, the third party would be also a processor, because there is a contract with one participating NSI. However, there are no specifications in the contract on how to process the data.
4. Third party involvement is neither regulated nor tailored: Under the fourth option fall all services provided by Google or by large physical activity tracker vendors such as Fitbit and Garmin. A further subtle, but important, distinction lies in the actual location of data processing, i.e. outside or inside the EU.

Third party involvement for ‘perfect data’ does not seem to be problematic for options 2 (regulation and tailoring) and 3 (only regulation) as there is a contract and respondents know what kind of data are submitted. This holds, perhaps, only under the condition that the third party is located in the EU. Option 4 always seems problematic, simply because there is no regulation.

* 1. TYPE OF DATA PROCESSING

The extent to which new forms of data are handled in-house is the second main new dimensions, which is driven by the desire that data transfer is minimized as much as possible to specified information needs.

Survey data collection has always been a subtle trade-off between respondent burden and measurement error. An all-knowing and fully motivated respondent could be interviewed with a minimal set of questions and a maximal respondent effort in deriving information. Literature distinguishes four main steps in answering a question: interpretation, information retrieval, judgment and reporting. An ideal respondent would go through these steps without any problem. Ideal respondents are rare and measurement errors occur because of a mix of inability and lack of motivation. Survey design, therefore, seeks to minimize the risk of such errors. Surveys have always collected a surplus of information from respondents in order to remove burden for respondents and to avoid measurement errors due to inability. An example of burden reduction is the household budget survey where respondents, through a time diary, submit all products they buy at a detailed classification level. The level of detail in both time of purchase and type of purchase is much higher than needed for the formal COICOP classification. This is simply because respondents would otherwise have to make a large series of computations and complicated judgments. An example of avoidance of error due to respondent inability are large series of questions about health, e.g. in the European Health Interview Survey (EHIS) and national versions of it, and about living conditions in housing surveys and EU-SILC. Simply asking a respondent how healthy he/she really is and lives in terms of exercise and nutrition would lead to measurement errors that render data useless. However, while there is an excess of data, respondents do know the individual answers. They may, however, not know or deduce the general picture that is extracted from their answers about, for example, their expenditure profile, health or living conditions

Even with careful design, answers to survey questions are susceptible to measurement errors, either consciously or unconsciously introduced by respondents. Unlike business surveys, there is no tradition in social/household surveys to interact with respondents on the quality of data. Respondents usually are allowed to provide do-not-know or refuse-to-say answers, to give inconsistent answers or answers that carry little information.

Smart surveys are motivated by the need to further remove respondent burden and to reduce measurement error. They do so through sensor measurements and/or data donation. These types of data may even provide proxies that are much closer to the concepts of interest than do questions. These surveys still follow the tradition of asking more information than needed to improve data quality, but with two new elements:

* Data may be collected that are unknown to respondents themselves
* A clear need arises to control quality of sensor/donated data.

Consider the earlier examples from the respondent perspective: Receipt scanning and processing in the household budget survey groups purchases into a profile that the respondent may not be aware of. Physical activity tracking gives direct estimates of types and intensities of activity that respondents only know approximately. Indoor climate measurements provide information that most households do not know at all.

These data collections still conform to the original information needs and the all-knowing respondent may have provided them. The need for quality control comes from the NSI perspective (i.e.: receipt scanning may give errors in text extraction or interpretation of what types of products texts refer to. Activity trackers show deficiencies in recognizing different types of transport modes, say an e-bike, or activities, say vacuum-cleaning the house. Indoor climate sensor systems use relatively cheap air quality sensors and for example measurement of particulate matter is relatively inaccurate). Aggregating data locally at a personal device/environment, would introduce measurement errors. Given that respondents may not know these data themselves, they can only be of partial help in adjusting errors. A fully motivated respondent would do this task, but many respondents would only do this to a certain extent. For an NSI it is, therefore, imperative to get data quality information.

Therefore, two events should be recognized in terms of data minimization settings, first, respondents do not know the data themselves and, second, sensor/donated data are subject to influential errors. With this in mind, again four options may be recognized:

1. Data are only mildly subjected to error and respondents are knowledgeable
2. Data are only mildly subjected to error but respondents can be of little assistance
3. Data are subjected to error but respondents are knowledgeable
4. Data are subjected to error and respondents can be of little assistance

Under settings 2 and 4, error handling would imply that the NSI includes data cleaning and processing rules in the respondent application, which would often mean additional data/information needs to be send to respondent devices.

* 1. OPEN DECISIONS IN SMART SURVEY DESIGN FROM THE LEGAL PERSPECTIVE

Crossing the two dimensions, third party involvement and type of data, we would arrive at 16 (= 4 x 4) possible scenarios. Let us first consider them separately where third party involvement is viewed from error-free data and knowledgeable respondents, and, vice versa, data minimization from no third party involvement.

Now let us cross the two dimensions. Table 1 shows the 16 options in a matrix. The colours give a recommendation on the desirability of the setting given the threats of unwanted access by third parties. The texts in the cells give recommendations on the type of processing: in-device, in-house or a mix of both.

If no third party (first row) is involved then the scenario is always allowed, while if there is no contract and no form of tailoring (last row) then it never is allowed. For the other scenarios it depends on the size of errors and the knowledge of respondents.

The recommended type of processing depends on the one hand on amount of errors and on the other hand on the tailoring of data collection. When errors are large, in-house processing is recommended, or at best a mix. Under contract and with tailored data collection (second row), the recommended type of processing does not change relative to no third party (first row). However, under large errors it becomes an open question whether it can be done under contract. When there is a contract but no tailoring, then only under modest errors and knowledgeable respondents could it be local processing; otherwise it must be in-house.

*Table 1: Proposed setting for the scenarios under third party involvement and type of sensor/donated data. Red cells indicate that the scenario is not allowed, orange cells indicate that there is a risk or doubt whether it can be allowed, and green cells indicate that the scenario is allowed. In text are given three processing options, namely in-device, unclear/mix, and in-house. ‘NA’ stands for ‘Not applicable’.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Third party | Type of sensor/donated data | | | |
| Modest errors  Assistance | Modest errors  No Assistance | Large errors  assistance | Large errors  No assistance |
| No third party | In-device | In-device | Mix | In-house |
| Contract Tailoring | In-device | In-device | Mix | In-house |
| Contract  No tailoring | Mix | In-house | In-house | NA |
| No contract  No tailoring | NA | NA | NA | NA |

To fix thoughts for table 1 recommendations, consider the case of activity tracking through wearable motion sensors. This could be in a health or lifestyle survey such as the European Health Interview Survey (EHIS). Activity tracking can be done with a range of devices from strongly commercial, customer market to research-grade and academic market. The customer-market devices do not allow for tailoring of methodology and data collection. Data are stylized through pre-trained but unknown routines. Research-grade devices come in various forms from raw data that can be extracted without intervention from the vendor and open source methodology to more restricted access to data and processing of data. Thus, all options exist from no third party to no contract and no tailoring. In activity tracking, the interest lies in two types of statistics: intensity of physical activity and type of physical activity. Intensity is relatively easy to measure through trackers, but mostly unknown to respondents unless it concerns specific sports. Type of activity is hard to measure through trackers, but relatively easy for respondents to classify unless it concerns very common activities such as sedentary behaviour. Hence, also in errors and knowledge all scenarios occur.

Table 1 tells us that working with research-grade trackers from which raw data can be extracted directly should be allowed, while working with larger customer-market devices should never be allowed. It also tells us that for estimating type of activity, the type of processing will largely be in-house, while for intensity it may be done in-device.

* 1. NEXT STEPS FROM THE LEGAL PERSPECTIVE

The working group produced the following recommendations for the near future:

* The working group should remain active and make use of the experience and knowledge that have been gained during the ESSnet. It could have an advisory role in follow-up projects;
* A joint and simultaneous application should be launched to national and European legal authorities asking for guidance and advice on the above issues;
* The working group focus should be extended to ethical considerations , and perhaps even NSI policy reasons, that may overrule legal constraints. In the meantime, it should consult parties that have gone through or are going through similar exercises
* A cross-national (ESS-wide) survey on trust in smart features should provide insight into perceptions and, consequently, boundaries to smart survey implementation. Two such studies have been conducted in NL already, which may be used as a starting point. Ideally, such a study should be linked to realistic and legitimate smart survey applications;

1. THE WP3 FRAMEWORK AT WORK
   1. WP3 GSBPM STAGES AND BUSINESS LAYER

In this section three questions are asked for WP2 case studies:

* What GSBPM stages have been investigated?
* What GSBPM sub-stages are particularly complex and/or time-consuming to design for smart surveys?
* What GSBPM sub-stages have become mature?
  + 1. What GSBPM stages have been investigated?

WP3 gives a detailed account of the GSBPM stages in smart surveys, e.g. deliverable 3.3. Out of the main stages (Identify needs, Design and build, Collect, Process, Analyse, and Disseminate), especially the first four (Identify needs, Design and build and Collect, Collect, Process) have been evaluated and elaborated. This is no surprise as WP2 case studies were mainly active in developing the processes between these stages.

At the start of the ESSnet, WP2 labelled its smart survey design activities as phase 0 (conceptual), phase 1 (collecting basic ingredients/components), phase 2 (functional and usability tests for methodology), phase 2’ (functional tests for IT and logistics), and phase 3 (field tests for choosing between the most promising designs). These phases can be aligned with the GSBPM stages and WP3 elaborations, but it must be realized that the WP2 phases are cyclical and not linear. The cyclical nature is twofold. The first cycle conforms to the plan-do-check-act (PDCA) cycle where, in the design of smart surveys, activities may alternate through the phases multiple times depending on growing experience and empirical results coming from tests. The second cycle consists of gradually ‘smartifying’ a survey, i.e. smart features are not added in one single go but gradually. Smart surveys grow in maturity depending on a complex business case evaluation in which response rates, data quality, costs and ease of implementation play a prominent role. Summarizing, smart survey implementations are not frozen and, consequently, also the GSBPM stages are revised and made more mature on a continuous basis.

The WP2 phase 0 concerned the GSBPM stage Identify needs. While starting from existing (ESS) applications (HBS for consumption, HETUS and Passenger mobility/Travel surveys for time use, EHIS and Lifestyle/Leisure surveys for physical activity, and EU-SILC and Housing surveys for indoor climate), both the representation and measurement arms of survey error had to be revisited. Representation of those persons/households that can apply smart features is different from those that belong to the target population. However, more importantly, smart features change the concepts that are measured. In fact, the main motivation for going smart is to get better proxies of concepts of interest and/or to assist respondents in automating tasks that are subject to measurement error. For the case studies on consumption and time use this GSBPM stage had been explored already quite extensively at the start of the ESSnet, but for case studies physical activity and indoor climate this certainly was not true.

The ESS dimension in this GSBPM stage was, of course, paramount. Within case studies consumption and time use considerable effort went into finding the largest common denominators. The large number of participating countries was very helpful for this purpose, e.g. see deliverables 2.2 and 2.5. Within the ESSnet, the two case studies, originating from existing ESTAT-funded projects, went back to the Identify needs stage.

WP2 phase 1 overlapped largely with GSBPM stage Design and build, but also implied going back to the Identify needs stage. The latter occurred when potential solutions did not work out, did not satisfy legal requirements and/or demanded for a more in-depth specification of the needs. Especially, for case studies physical activity and indoor climate, it was unclear at the start what was and could be being measured by sensor systems and, in particular, what was the accuracy of the measurements. Furthermore, sensor systems are almost without exception produced by commercial vendors and the boundaries of third party involvement had not yet been determined.

WP2 phases 2 and 2’ overlap with the GSBPM stages Design and build and Collect, where phase 2 focusses on processes and phase 2’ on applications. Given the cyclical nature of the case studies, the GSBPM Collect stage was only evaluated in experimental, small-scale data collections going for usability, user experience and first impressions of respondent engagement. In terms of methodology and logistics, the main concerns were with app user interfaces, recruitment material, respondent instructions, interviewer instructions, handling of external devices, and properly performing smart tasks. In terms of infrastructure, it was about stress tests, robustness of external devices, security checks, improving configurability across countries, and back-office monitoring and case management.

Finally, phase 3 ventured into the GSBPM Collect and Process stages, although field test results also imply that the Design and build stage will need to be revisited. A relatively large-scale data collection was conducted which was presented to respondents as a regular, production survey. Furthermore, helpdesks were installed, monitoring and analysis and case management were performed and resulting data were processed up to the point where they could be compared to regular statistics.

So which stages were explored for each specific case study? Tables 2 to 4 consider the four GSBPM stages Identify needs to Collect for each case study. Case study consumption has been evaluated and elaborated furthest as it also had a field work component from which HBS statistics were derived. Case studies physical activity and indoor climate were limited to collection of sensor data and did not consider additional contextual data or questionnaires. Case study physical activity also employed an ad hoc, mostly manual, data collection strategy during small-scale tests. Case study indoor climate did use the data collection processes of the commercial vendor. These need, however, be adapted in a next step, since passing data through a third party without control over data crosses legal boundaries.

*Table 2: GSBPM Identify needs substages and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Data providers | TRUE | TRUE | TRUE | TRUE |
| Sensors | TRUE | TRUE | TRUE | TRUE |
| Incentive schemes | TRUE | FALSE | TRUE | FALSE |
| Security & confidentiality | TRUE | TRUE | TRUE | FALSE |
| Data storage | TRUE | TRUE | FALSE | TRUE |
| Data processing | TRUE | TRUE | TRUE | TRUE |
| Consent | TRUE | FALSE | TRUE | TRUE |
| Privacy | TRUE | TRUE | FALSE | FALSE |

*Table 3: GSBPM Design and build substages and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Data collection tools | TRUE | TRUE | TRUE | TRUE |
| PPT | TRUE | TRUE | FALSE | FALSE |
| Monitoring system | TRUE | TRUE | FALSE | TRUE |
| Metadata management | FALSE | TRUE | FALSE | FALSE |
| Processing | TRUE | FALSE | TRUE | TRUE |
| Deployment production | TRUE | TRUE | FALSE | FALSE |

*Table 4: GSBPM Collect substages and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Data collection strategy | TRUE | FALSE | TRUE | FALSE |
| Security-privacy issues | TRUE | TRUE | FALSE | FALSE |
| Questionnaire | TRUE | TRUE | FALSE | FALSE |
| Data collection tool | TRUE | TRUE | FALSE | FALSE |
| Data acquisition | TRUE | TRUE | TRUE | TRUE |
| Data storage | TRUE | TRUE | FALSE | FALSE |
| Data preparation | TRUE | FALSE | TRUE | TRUE |
| ML techniques | TRUE | FALSE | TRUE | TRUE |
| Standardization | FALSE | FALSE | FALSE | FALSE |
| Encoding | TRUE | FALSE | FALSE | FALSE |
| Derivation | TRUE | FALSE | FALSE | FALSE |
| Linkage and enriching | TRUE | FALSE | FALSE | FALSE |

WP3 did not elaborate the GSBPM stage Process, but part of this stage was explored in case study consumption. Sub-stages under Process are outlier detection and editing, selection of (sufficiently) complete households, linkage of auxiliary variables for weighting, aggregation of household purchases over the reporting period and over household members, adjustment of sample inclusion weights, and adjusting/weighting for nonresponse. HBS survey data have been transformed and aggregated and weighted for nonresponse in a way similar to the regular HBS. These steps differ from the regular survey in two ways. The first is that paradata on in-app navigation behaviour is available and can be used as information to decide whether a household can be selected as complete. The second, much more influential, is that ML classification produces statistical predictions of COICOP and no absolute values. For now, COICOP predictions have been included as weights, i.e. one product is assigned to multiple COICOP categories with weights proportional to classification probabilities.

* + 1. What GSBPM sub-stages are particularly complex and/or time-consuming?

The previous section gave a global account of the sub-stages that were considered, but not on how intensive they were in terms of design, development and logistics preparations. Retrospectively, it can be concluded that the most intensive stages are:

* Identify needs
  + Choice of external sensor systems: For physical activity and indoor climate extensive explorations were performed of candidate sensors, including lab tests;
  + Third party involvement in data storage and data transfer: Legal boundaries to third party involvement had not been considered prior to the ESSnet and played a very dominant role in the time use case study and to a lesser extent in the external sensor system studies;
  + In-app navigation paradata design: At the start of the ESSnet there was no experience or literature on how to use paradata about how respondents use applications;
  + Smart feature accuracy requirements: Sensor measurements imply a change to measurement concepts. While traditional measurements through questions are subject to measurement error, the need for quantifying sensor error by stakeholders is more explicit.
* Design and build
  + User interface for respondent interaction of active smart features: Thus is a broad topic and ranges from HBS product lists, in-app editing of text extraction to geo-location visualization. It is a complex and fine balance what to ask from respondents and how. This aspect of smart surveys requires several rounds of testing to converge to a solution;
  + External sensor system data preparations: External sensor system data need to be linked to NSI systems. This requires separate channels;
  + External sensor system logistics (case management, monitoring): In the ESSnet case studies, sensor systems were provided to respondents. The logistics are mostly new to NSI’s;
  + Paradata monitoring: Paradata on respondent in-app behaviour need to be monitored through adequate scripts and queries;
  + Machine learning model preparations: Smart surveys use new types of data and push part of the processing to the respondent devices. However, also in-house processing is mostly new;
  + Online/active learning of existing ML models: Smart survey data may be subject to strong dynamics in time, so that frequent updating and retraining are imperative to keep performance at acceptable levels;
  + Sensor measurement IT: The correct handling of in-device sensors such as keeping location tracking running on the background is extensive work due to the variety of OS and models;
* Collect
  + Interviewer instructions and training: Interviewers may play an important role in improving respondent engagement but need training and get acquainted with tools themselves;
  + Deployment preparations: Smart surveys are typically longitudinal and set different specifications to case management, monitoring and analysis. AT NL, smart surveys are seen as a separate data collection channel, next to web, CATI and CAPI;
* Process:
  + Deployment preparations: While smart survey data may resemble non-smart data, they really are not and require a complete revision of the post-survey processing;

Some of the substages will get less extensive with time as experience grows. The most demanding substages for new smart surveys or existing smart surveys that add new smart features are:

* User interface for respondent interaction of active smart features;
* Machine learning model preparations
* External sensor system logistics (case management, monitoring)
* Deployment preparations
  + 1. What GSBPM sub-stages have become mature?

As was mentioned previously, the ESSnet case studies went through various design and development iterations during 2020 and 2021. This has especially led to crucial experience in the relatively early GSBPM stages Identify needs and Design and build. The simultaneous development of the physical activity and indoor climate case studies was very useful to learn how to embed external sensor systems in existing surveys. One such case study would have been anecdotal. The consumption case study has brought crucial knowledge about respondent engagement and respondent interaction.

Given the four case studies and the additional activities in the ESSnet working group legal, see section 4, it can be concluded that the GSBPM Identify needs stage has become mature. For a new smart survey or a new smart feature in an existing survey, it would be relatively straightforward how to collect and evaluate the basic components of a smart survey. There still is an open question about the legal boundaries for in-device versus in-house processing, but legal departments have been heavily involved and such discussions would be much more concrete than at the start of the ESSnet.

For the GSBPM Design and build stage, smart surveys have become mature when restricting attention to two of the smart features: Local storage and processing, and Employment of in-device sensors. In case studies consumption and time use, both features have been explored in-depth and incorporated in design in all possible ways. For the smart feature Employment of external sensor systems, there still are various steps to go. ESSnet case studies did not link those to existing systems, nor did they embed them in an existing survey with additional context questions. The smart feature Linkage to public online data has been explored in case study time use, but only on small-scale. The smart features Data donation through the respondent and Data donation through the NSI have not been explored at all.

The GSBPM Collect stage has been more or less completed for the consumption case study with recommendations on the data collection strategy and respondent interaction strategy.

Summarizing:

* The Identify needs stage can build on extensive ESSnet experience;
* The Design and build stage has been far advanced for smart surveys with internal sensors;
* The Collect stage can borrow experience from the consumption pilot for smart surveys with internal sensors;
  1. WP3 APPLICATION LAYER AND BUILDING BLOCKS

WP3 paid also extensive attention to the application building blocks through its application layer. WP3 groups blocks under four main categories and with Metadata management as a fifth overarching block. The four blocks are Smart data tools, Smart survey monitoring, Smart data acquisition, and Smart data processing.

In this section, two questions are asked for WP2 case studies:

* What building blocks have been developed?
* To what extent can building blocks be used within, or even across, smart survey domains?
  + 1. What building blocks have been developed?

At the start of the ESSnet, none of the WP2 case studies was in production. However, applications within case studies consumption and time use had been applied already on a quantitative scale. For consumption, the Household Budget Survey app existed at the start of the ESSnet, but without in-app text extraction and without machine learning routines for COICOP classification. The time use case study built upon the existing MOTUS application and platform that had been applied to many production settings, but without any smart features. The physical activity case study could make use of experience and small-scale tests with a variety of sensor systems prior to the ESSnet, so that the sensor selection and evaluation had a jumping start. For indoor climate, the case study started completely from scratch.

Tables 5 to 9 present the status for the application subblocks for each case study. For the case studies physical activity and indoor climate the number of completed blocks is modest and more work is needed. For consumption almost all blocks have been developed, except for the machine learning routines that need further optimization. For time use, there is a platform. However, the smart features have only been investigated at a basic level.

*Table 5: Smart data tools subblocks and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Frontend | TRUE | TRUE | FALSE | FALSE |
| Backend | TRUE | TRUE | FALSE | FALSE |
| Motion | NA | NA | TRUE | NA |
| Location | NA | TRUE | NA | NA |
| OCR/NLP | TRUE | NA | NA | NA |
| Air quality | NA | NA | NA | TRUE |

*Table 6: Smart survey monitoring subblocks and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Fieldwork monitoring | TRUE | TRUE | FALSE | FALSE |
| Respondent assistance | TRUE | TRUE | TRUE | TRUE |
| Staff training | TRUE | FALSE | FALSE | FALSE |
| Paradata analysis | TRUE | FALSE | NA | NA |
| Contextual data analysis | TRUE | FALSE | TRUE | TRUE |

*Table 7: Smart data acquisition subblocks and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Passive data | TRUE | TRUE | TRUE | TRUE |
| Active data | TRUE | TRUE | FALSE | FALSE |
| Data storage | TRUE | TRUE | FALSE | IN PART |

*Table 8: Smart data processing subblocks and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Analysis and validation | TRUE | IN PART | TRUE | TRUE |
| Preparations | TRUE | IN PART | IN PART | IN PART |
| ML routines | IN PART | IN PART | TRUE | IN PART |
| Linkage and enriching | TRUE | FALSE | FALSE | FALSE |
| Quality management | FALSE | TRUE | FALSE | FALSE |

*Table 9:Metadata management subblocks and the four case studies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Consumption | Time use | Phys activity | Indoor climate |
| Metadata management | TRUE | TRUE | FALSE | FALSE |

* + 1. To what extent can building blocks be re-used?

Important requisites of a smart survey solution are shareability across countries, in particular ESS, and modularity across survey domains. Deliverables 2.5 and 2.2 are devoted to these two features of smart surveys and conclusions are not repeated here. Nonetheless, it should be discussed what building blocks specific to smart features can actually be re-used. The smart features that have been considered in the ESSnet are Local storage and processing, Employment of internal smart device sensors, and Employment of external sensor systems. For the internal sensors it concerned camera and location (GPS, WiFi, GSM) and for external sensors it consisted of motion sensors and a range of air quality sensors.

The type of data collected with internal smart device sensors is very different. Even with the same sensor, the use of the sensor data can be hugely different. Consequently, machine learning routines are also tailored to each application. Two examples are given.

Camera images may serve a very wide range of purposes. In the ESSnet it concerned receipt scanning. However, in smart survey studies outside the ESSnet, many other applications have been investigated. While OCR is fairly generic, NLP methods for the consumption case study have been completely made fit to the layout and logic of receipts. They cannot be re-used for any other purpose.

Location tracking can be re-used, but derivations that are dependent on location tracking such as transport mode prediction, prediction of travel purpose and stop-track detection strongly depend on definitions and classifications used.

Hence, on a very basic level, namely letting a sensor produce data without any further interpretation, sensor measurements can be re-used. So microservices/building blocks for location tracking and camera images plus OCR can be re-used, but any further action needs to be developed individually. WP3 devoted some effort to a generalized machine learning module, but this seems infeasible or inefficient at best.

Re-use of a series of building blocks linked to a specific internal sensor, thus, is limited to the same or very similar purpose. Per purpose separate processing building blocks need to be prepared.

For external sensor systems, the potential of re-use is greater. These systems have been designed for a well-defined, specific purpose. They can, therefore, be linked to different surveys. However, their specific purpose is also a drawback; they allow for little or no tuning in measurements such as frequency or type of sensors. Given the specific purpose, often machine learning routines have been studied and reported in literature. It is feasible to create a set of machine learning routines that would serve multiple applications.

The application layer, i.e. the whole of building blocks suggested by WP3, obviously goes beyond mere software tools. It also consists of methodology and logistics. Here, much can be re-used. There are strong parallels for all categories that WP3 proposed. The app frontend and backed may be based on the same basic lay-out and structure, monitoring tools may consist of an adjustable dashboard, helpdesks and other supporting functions may be relatively easily informed to a new smart feature, etc. Hence, the re-use of smart survey building blocks lies mostly in the methodology and logistics behind the building blocks.

* 1. CLASSIFICATION OF WP2 CASE STUDIES THROUGH WP3 SCENARIOS

WP3 identified five characteristics of smart surveys that determine their implementation:

* Type of data acquisition (Active/Passive data), depending on the sensor used to collect data
* Type of data provider (Respondent/Third parties)
* Type of data storage/management (Local (on device)/ In-house (NSI)/Centralized (Platform) data storage)
* Type of data processing (Local (In-app)/ In-house (NSI)/Centralized (Platform) data processing)
* Type of service deployment (NSI/Centralized (Platform)).

The four WP2 case studies are classified based on these five characteristics. To follow-up on the considerations in section 4, also the nature of the data is added as a sixth characteristic:

* Nature of data (minor issues & knowledgeable respondents, minor issues and no knowledgeable respondents, major issues & knowledgeable respondents, major issues and no knowledgeable respondents)

Tables 2 to 5 present the classification of the four case studies according to the six characteristics. Smart features in use are:

* Product search and receipt scan processing for consumption;
* Geo-tracking/fencing for Time use
* Motion sensors plus optional heart rate for Health/Physical activity
* Indoor air quality sensors for Living conditions/Indoor climate;

*Table 2: Classification of the smart HBS.*

|  |  |  |
| --- | --- | --- |
| Property | Classification | Explanation |
| Data acquisition | Active | Respondents are involved directly in checking selecting relevant parts of receipts, editing OCR/NLP results and in providing context on unknown products. |
| Data provider | Respondent | There is no third party |
| Data storage/management | In-house | Both processed receipt data and receipt images are submitted to the backend |
| Data processing | In-app + in-house | Products are classified in-house. OCR/NLP can be done in-app, but for now in-house is available as backup. |
| Deployment | NSI | Services are either replicated through a docker container option or interoperable through specific NSI solutions. Machine learning routines should be coordinated centrally. |
| Nature of smart data | Major issues  Knowledgeable | Processing of receipts, either scanned or digital, is subject to errors in OCR and language processing. These steps may be further perfected, but is likely that for a subset of shops and store types assistance is needed.  Respondent are knowledgeable about receipts, but assistance can be burdensome (which is the motivation for going smart). |

*Table 2: Classification of the smart time use.*

|  |  |  |
| --- | --- | --- |
| Property | Classification | Explanation |
| Data acquisition | Passive | Geo-locations are collected passively and only offered to respondents to assist in completing time use diaries. Respondents can ignore but not change.  NB: When location data are used directly for travel surveys, then an active component is needed. |
| Data provider | Respondent | All data are provided by respondents.  NB: When using location data directly for travel surveys, then external points-of-interest data may be added. |
| Data storage/management | Local | Location data are not used themselves in producing time use statistics. |
| Data processing | In-app | All processing is local as it is merely to support respondents. |
| Deployment | Centralized | The MO:TUS platform in principle assumes shared services although replicated services have been investigated in innovation projects parallel to the ESSnet. |
| Nature of smart data | Minor issues  Not knowledgeable | Respondents know about the type of locations they have visited but usually do not know exact times at which they did and where exactly these locations were. Location data is subject to missing data and outliers/noise. However, as data are used only as extra information, there are no major issues.  NB: When location data would be used directly in travel surveys then major issues are expected. |

*Table 2: Classification of the smart physical activity.*

|  |  |  |
| --- | --- | --- |
| Property | Classification | Explanation |
| Data acquisition | Mostly passive | Activity trackers in principle are passive measurement devices. It is also not recommended to let respondents revise intensity of activity, but they may be involved in revising the predicted type of activity. |
| Data provider | Respondent | Assuming research-grade trackers, there is no third party involved and data are retrieved directly from the devices. |
| Data storage/management | In-house | Assuming research-grade trackers, there is no third party involved and data are stored directly on internal in-house databases. |
| Data processing | In-house | Activity trackers do not allow for local processing and all processing is done afterwards. |
| Deployment | NSI | Assuming research-grade trackers, the trackers are provided by the NIS itself. |
| Nature of smart data | Minor/major issues  Not knowledgeable | For type of activity, respondents are knowledgeable and activity tracker predictions do have issues when classification becomes detailed.  For intensity of activity, respondents are not knowledgeable, but trackers provide accurate data. |

*Table 5: Classification of the smart indoor climate.*

|  |  |  |
| --- | --- | --- |
| Property | Classification | Explanation |
| Data acquisition | Mostly passive | Air quality systems in principle are passive measurement devices. It is also not recommended to let respondents revise measurements, but they may be involved in providing context to unexpected/implausible values. |
| Data provider | Third party | In the ESSnet only third party systems have been evaluated that do not allow to retrieve data directly. However, portable research-grade systems do exist with a much more limited range of measurements. |
| Data storage/management | In-house | Data are stored in-house, but may go through a third party storage environment. It still has to be investigated whether such by-passes can be avoided. |
| Data processing | In-house | All data are processed afterwards. This applies mostly to sensor fusion, i.e. the aggregation of individual air quality parameters to overall profiles, and to understanding the context of measurements, i.e. the household life style and the dwelling characteristics. Data of third party systems typically already went through a number of pre-processing steps. |
| Deployment | NSI and third party | The air quality systems are provided and distributed by the NSI itself. However, without a bypass of the data flows, part of the process is in the hands of the third party. This is not an acceptable solution given the discussion on legal requirements. Ultimately, a direct data flow is imperative. |
| Nature of smart data | Minor issues  Not knowledgeable | Air quality parameters, such as particulate matter, TVOC and CO2, vary in the complexity and accuracy of measurement. For the purpose of general population surveys, the set of air quality estimates is sufficient. Respondents cannot be consulted as they have no knowledge of air quality. |

1. NEXT STEPS AND RECOMMENDATIONS

ESSnet Smart surveys has been a unique project. It ventured deep into survey data collection design and analysis. Especially, the wide range of usability tests, functional tests and field test are unprecedented. These have brought valuable knowledge in an international context. However, it also is clear that there is a friction between expectations at the start and results at the end of the ESSnet. Uncertainties about what can and cannot be done in the legal domain and COVID-19 both severely impacted the project. But even without these influences, it is likely that there would have been some friction. Smart surveys require multiple development cycles in order to create a baseline architecture and infrastructure. During the ESSnet the four pilot studies indeed went through multiple cycles and each application now is much more nature than it was at the start. For the counterpart project WP3, this meant they had to be patient and had to observe accumulating knowledge. These efforts have not been in vain. The enhanced framework produced by WP3 proves to be a valuable stepping stone and line of thinking along which smart surveys can be further elaborated and made more mature. In this end report signs of this emerging utility have been substantiated. Given that WP pilot studies now start to move to GSBPM stages Process, Analyze, Disseminate, it is also apparent why these still have been underexplored and underevaluated by WP3.

So what is next? ESSnet Smart surveys consisted of the first, but crucial, steps into development and implementation of smart surveys in official statistics. These steps are again unprecedented in official statistics and, even more generally, in general survey statistics. The ESSnet presentations and derived papers at other (inter)national meetings attracted a lot of interest universally. Findings in the WP2 pilots are broadly relevant and can be published in academic journals.

Official statistics have clear advantages over other survey organisations. They often have access to administrative data and other forms of existing data on a legal basis. They perform repeated surveys with overlapping themes so that there is time but also incentive to learn. They still often have a sizeable interviewer pool and resources to conduct surveys on population levels; lifting the importance and magnitude of smart survey applications. It is, thus, not surprising that smart survey innovation is initiated in NSI’s. This momentum should be kept for the coming years as more cycles are performed and more and more applications may become mature. Along with it the WP3 framework may become more and more elaborated and also move to later GSBPM stages.

To date, given the diversity in NSI case management systems, logistics and skills, it is unclear whether smart surveys can be (fully) replicated. However, a prominent feature that smart surveys share with smart statistics is the reliance on machine learning and other data science procedures. There is considerable incentive to replicate such procedures. They imply a large investment, so that there is efficiency to be gained. They are also not yet embedded in case management systems and NSI’s can basically start from scratch by sharing microservices.

The following recommendations are made:

* Continue to work on smart survey frameworks by continuing to observe how smart surveys go through development cycles and get mature;
* Explicitly embed the PDCA (plan-do-check-act) cycle in smart survey frameworks as survey climate and communication channels will continue to change;
* Search for clarification and guidelines on the most important outstanding legal discussion: how to handle quality data and contextual data on smart surveys from the GDPR data minimization viewpoint;
* Explore and expand smart features that have only been explored briefly such as linkage to publically available online data and different form of data donation;
* Harmonize and coordinate at least the machine learning methodologies behind training, retraining and predicting based on sensor data and other forms of new data;

1. https://www.motusresearch.io/ [↑](#footnote-ref-1)
2. https://torvub.be/ [↑](#footnote-ref-2)
3. These percentages are based on raw data; weighting will increase them further. [↑](#footnote-ref-3)