ESSnet Smart Surveys: Smart features, machine learning and privacy

Annemieke Luiten (Stat NL) & Barry Schouten (Stat NL & UU)

MASS22, June 16-17, 2022

ESSNET SMART SURVEYS

Smart surveys?

Smart surveys include one or more of the following smart features:

- Local/in-device storage and processing
- Employment of internal mobile device sensors
- Employment of external sensor systems
- Linkage to public online data
- Data donation through the respondent
- Data donation through the statistical institute

In other words, smart surveys bridge the gap between surveys and big data, while keeping respondents at the centre of data collection

ESSnet Smart surveys

ESTAT funded project running from Jan 2020 to June 2022.
A six month extension halfway due to COVID-19 impact on fieldwork and legal discussions

Two main work packages:

- Smart survey case studies: Consumption, Time use, Health and Living conditions
- Smart survey architecture and infrastructure

Focus on shareability across ESS countries

Between December 2020 and March 2022 a working group considered smart surveys in the context of GDPR and ethics.

Smart surveys

Smart surveys employ features of smart devices in order to:

- Reduce respondent burden;
- Automate measurement-error-prone tasks;
- Improve survey experience;

Goal remains to collect data that fully and only serve the concepts of interest in the information need, BUT

- Automation is not perfect, i.e. sensor/donated data contain errors;
- Survey institute needs to know context;

EXAMPLE: Travel survey

Need: Where, how and why do respondents travel

Location tracking: Gaps, outliers, incomplete on transport mode and purpose

Active-passive smart data collection

Active data collection = Respondents are involved in interpretation of the sensor task, retrieving information through the sensor task, judging the sensor data, and/or submitting the sensor data.

Why active data collection?

1. Respondent engagement: To increase respondent control, to make the survey more enjoyable, to feedback insights; ↓↑
2. Sensor error adjustment: To adjust for errors that may occur in collecting sensor data; ↑
3. Legal (ethical): To conform to data minimisation principles in data collection legislation (such as GDPR); ↓

Motivations 1-3 are conflicting

WG legal smart survey presumptions

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, 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.
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.

Two main dimensions in legal discussions

1. The extent to which third parties involvement is regulated, because access to personal data is limited to those that have a legal mandate for data collection for statistical purposes;

2. The extent to which new forms of data are handled in-house, providing that data is minimized as much as possible to specified information needs;

Third party involvement

1. There is no third party involvement
2. Third party involvement is both regulated and tailored to the specific need: The third party is a processor and intervenes “on behalf” of the NSI that signed the contract and stated the specific need.
3. Third party involvement is regulated but not tailored: The third party is also a processor, because there is a contract with one participating NSI. However, there are no specifications on how to process the data.
4. Third party involvement is neither regulated nor tailored: All services provided by Apple/Google or by large physical activity tracker vendors such as Fitbit and Garmin. A subtle, but important, distinction lies in the actual location of data processing, i.e. outside or inside the EU.

Processing of smart survey data

Classification of new forms of data and errors that may occur:

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

Example: Physical activity tracking

Third party involvement

1. Research-grade trackers allowing for direct retrieval of raw data;
2. As 1, but using existing vendor ML models to classify;
3. As 2, but using also the vendor online services ‘as is’;
4. Customer-market trackers such as Fitbit, Garmin;

Data classification:

1. Estimating active versus inactive time periods;
2. Estimating specific intensity of activity according to fit norms;
3. Predicting general types of activity;
4. Predicting sedentary behavior;

Classification of possible scenarios

Red = not allowed, orange = risk or doubt, green = allowed.
Processing options: in-device, unclear/mix, and in-house. NA =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

WG conclusions

Smart surveys are challenging in terms of GDPR:

- They use personal devices;
- They collect new forms of data with new types of errors and data are partly unknown to respondents themselves;

It is as yet unclear how to deal with quality metadata and handling of errors in donated/sensor data from the perspective of data minimization.

There are differences in how strict GDPR criteria are interpreted by countries/NSI’s.

Future

- WG will remain active as a network and may become more active when follow-up projects are launched;
- Jointly and simultaneously launch an application to national authorities in multiple countries, or even EDPS;

- Extend to ethical/policy boundaries in a follow-up project;
- A cross-national 'survey' in which persons are asked about their opinions on privacy and trust in surveys employing smart functions.

- Interested in joining? Please contact jg.schouten@cbs.nl