CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection

10-12 October 2018, Geneva, Switzerland

Session 1

21 September 2018

A Case Management System for Social Surveys based on the Business Process Model and Notation (BPMN) standard: Implementation and experiences at Statistics Austria

Josef Kytir (Statistics Austria)

Josef.Kytir@statistik.gv.at

Abstract

Mixed-mode data collections for social surveys could become very complex in respect to case management. This complexity is especially challenging when combining automated with individual case management, reacting on both at any time during the course of the data collection, individual preferences of respondents as well as of interviewers in case of surveys including a CAPI data collection episode.

In 2014, Statistics Austria started developing a new technical survey infrastructure (STATsurv) focussing on social surveys. From the beginning of 2018, all social surveys including the Labour Force Survey and EU SILC are successfully running under the new Data Collection System. For designing and running survey workflows the global standard for process modelling, the Business Process Model and Notation (BPMN) is used. BPMN is a graphical representation for specifying business processes. Non-IT staff of a National Statistical Institute can use an open source modeller to design graphical survey workflows according to BPMN standards. The BPMN specification also includes extensions utilized by IT staff for adding all the information necessary for the technical execution of the graphically designed workflow. This results in an XML specified workflow ready for execution by the use of a workflow and decision automation platform like Camunda BPM.

The paper deals with the way Statistics Austria modularises the collection phase of surveys and how these modules help designing and running survey workflows. Examples of how individual case management is feasible within automated and standardized workflows are presented.

A Case Management System for Social Surveys based on the Business Process Model and Notation (BPMN) standard: Implementation and experiences at Statistics Austria

Josef KYTIR, Statistics Austria

Although administrative data and registers are used more and more, sample survey data remain an important pillar in social statistics. In the European Statistical System a new framework regulation (Integrated European Social Statistics – IESS) aims to streamline seven sample surveys for persons and households needed for the production of European Social Statistics. As in other areas of statistical production, the technical, organisational and social environment for survey statistics changed a lot over the last decades. Quality standards increase whereas a considerable decrease in participation rates could be observed in many countries. One strategy to deal with these challenges is to use more complex survey designs, including mixed-mode designs, sophisticated contact strategies and adaptive approaches. However, combining these strategies and approaches in a single survey results in rather complex business processes during the data collection of a survey. Running several such surveys simultaneously poses a considerable burden on data collection units in Statistical Offices.

Of course, the most essential element of a sample survey is the questionnaire. During the last decades, many IT-tools for designing and running electronic questionnaires were developed. For simple and straightforward survey designs a questionnaire and some basic tools for monitoring completes and refusals of the sampled respondents would be sufficient. However, as survey designs get more and more sophisticated automated case management systems become essential to run surveys in an efficient and cost effective way. In addition to efficiency and cost effectiveness National Statistical Institutes have to consider other quality dimension too. Therefore, you end up with the demand for a tool combing as far as possible the automatisation of the data collection process with the feasibility for non-automated, individual case management. This non-automated case management should allow fieldwork management staff to overrule or complement the automated processes to be able to react on individual preferences of respondents as well as of interviewer in case of surveys including CAPI as a data collection mode.

Much time and effort was used to develop electronic tools for designing complex survey questionnaires. Despite the growing importance there is obviously much less attention paid to develop adequate solutions for case management systems suited for running complex statistical sample surveys. In 2017, Eurostat launched a grant titled MIMOD (Mixed Mode Designs for Social Surveys) that aims for a cooperation in tackling challenges related to data collection for mixed-mode surveys. One of its work packages investigates the landscape of case management systems in use at the National Statistic Offices within the European Statistical System. First results show that there are very different approaches to the

architecture of such case management systems, sometimes even when requirements are very similar.

How to standardise and modularise the processes during fieldwork/data collection?

In the Generic Statistical Business Process Model (GSBPM; see https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0) all activities related to the implementation of a statistical survey are described in the "Collect" phase. This phase is split up into four sub-processes. The actual fieldwork of a survey relates to the third sub-process labelled "run collection" which is described in the following way:

> "This sub-process is where the collection is implemented, with the different instruments being used to collect or gather the information, which may include raw micro-data or aggregates produced at the source, as well as any associated metadata. It includes the initial contact with providers and any subsequent follow-up or reminder actions. It may include manual data entry at the point of contact, or fieldwork management, depending on the source and collection mode. It records when and how providers were contacted, and whether they have responded. This sub-process also includes the management of the providers involved in the current collection, ensuring that the relationship between the statistical organisation and data providers remains positive, and recording and responding to comments, queries and complaints. …"

Although most of the essential elements are addressed, the description does not help to operationalise the activities related to survey fieldwork in a way that an automated case management system could easily be implemented. Therefore, at first it is necessary to reflect about several conceptual questions, the two most important ones being:

1. How to structure the relevant processes involved in survey data collection in a way that is as simple and generic as possible?
2. What activities in which situation are appropriate being controlled by an automated case management system and for what kind of activities this is not feasible or useless?

Each automatisation of processes is always closely linked to a standardisation and modularisation of tasks. When applying these concepts to the "run collection" process the most important building block or module of the whole process could be entitled as the "Data Collection Episode" (DCE). We define the essential elements of such a Data Collection Episode as follows:

- The episode addresses a single sample unit.
- The episode has to have a pre-defined start date and a pre-defined length or end date.
- The episode has to have a pre-defined mode (CAPI, CATI, CAWI …).

- Entering the questionnaire of the survey by the respondent or by the interviewer is allowed only in the time period between start date and the end date of the episode (or in a pre-defined part of that period).
- After termination, an episode has to have an assigned outcome category, at last indicating the result of all attempts within the episode to come up with a completed questionnaire in the given mode.
- The termination of an episode could be done in two different ways:
    - Active termination - The episode is closed by human interaction within the pre-defined time span, indicating the outcome (response, non-response, non-eligible …) and eventually providing all the other necessary information and comments on the sample unit. For interviewer-based modes, usually most of the episodes are closed actively.
    - Passive termination – No active closing happened until the predefined end of the episode. Consequently, the episode has to be terminated automatically. A predefined algorithm based on information available decides on the outcome code of the episode. Of course, the most important information in that respect is always the status of the questionnaire. For a self-administered mode like CAWI most of the episodes are subject to passive termination, most likely resulting in a refusal (respondent never entered the questionnaire).

For any survey design, the “run collection” sub-process for a single sample unit could now be modelled as a pre-defined sequence of 1 to n data collection episodes using the same or different survey modes. By default, the outcome of the chronologically last episode indicates the final outcome category (using a standard like the Lynn et al (2001) ISER working paper on Recommended Standard Final Outcome Categories and Standard Definitions of Response Rate for Social Surveys) of the whole “run collection” sub-process for that sample unit.

A second element common to most sample surveys is sending personalised information (letter, e-mail, SMS) to a sample unit. This includes prenotation letters, reminder mails, thank you postcards et cetera. For an automated fieldwork process it is sufficient to create a task referring to a pre-defined template. The only function of this Communication Request Task is triggering the production of a document or an e-mail according to the specified template. The production of the document itself or the sending of the e-mail or SMS could then be managed in separate business processes.

Data Collection Episodes and Communication Request Tasks are the two basic building blocks of the “run collection” sub-process of any survey. However, to model this process ready for automated execution you need some additional features. These features are a start event, various types of gateways to model different paths through the process, intermediate events waiting for some piece of information (eg. triggered by human interaction in a GUI), timer events controlling for dates and durations, and an end event for the whole workflow. Combining all these elements according to the design and the contact strategies of a particular survey results in a default fieldwork workflow for this survey.

Automatically each sample unit is now put to its default fieldwork workflow at a pre-defined date based on information provided in the sample data and an algorithm defined in the start

event and pass through to the end event of the workflow. As a result, when reaching the end event, a sample unit has a final outcome category indicating what was finally achieved during the fieldwork phase, data from the questionnaire (statistical data) depending on questionnaire status, and all kind of process data (paradata). The sample unit is now ready to proceed to the next phase of the GSBPM.

As an (simplified) example, a default fieldwork workflow for a mixed-mode survey having individuals as sample units should met the following requirements:

1. For all sample units fieldwork should start on Monday, 15th of October 2018.
2. Send a personalised prenotice letter to a sampled person amongst others providing a web link to a CAWI (Computer Assisted Web Interviewing) questionnaire.
3. Make the CAWI questionnaire available to the sample unit.
4. If the CAWI questionnaire is not completed within 3 days send a first reminder letter to the sampled person.
5. A personal email address could be provided in the CAWI questionnaire. In case of completing the questionnaire, send a thank you email to the respondent. If no email address is available send a thank you postcard.
6. Disable the availability to a non-completed questionnaire after 8 days.
7. In case of not completing the CAWI questionnaire send a second reminder to the sampled person amongst others indicating that a field interviewer will contact him or her.
8. 3 days after triggering the second reminder, the field interviewer should start contacting the sampled person and trying to carry out a Computer Assisted Personal Interview (CAPI).
9. The field interviewer has a maximum of 15 days to finish his work.
10. Send a thank you postcard (or email if possible) in case of a CAPI completed questionnaire.

Figure 1 shows a graphical representation of the survey fieldwork workflow meeting these requirements. Symbols and wording follow the global standard for process modelling, the Business Process Model and Notation (BPMN). The BPMN based workflow graphic is created by using the Camunda Modeler, an open source desktop app for editing BPMN process diagrams. The app is available at the website of Camunda (https://camunda.com/products/modeler/).

In principle, the Camunda modeler app will allow non-IT staff members in a National Statistical Institute to produce a graph of the fieldwork workflow according to his or her survey design and contact strategies using the two basic building blocks (Data Collection Episode and Communication Request Task) and the other features described above. In practice, it is necessary to collaborate with an IT staff member familiar with the necessary technical specifications of these elements and several additional features already implemented, for example checking for an existing interviewer assignment of the sample unit before starting a CAPI data collection episode (see Figure 1). As advanced design requirements could result in rather complex processes it will eventually be useful to involve someone with skills in simplifying and streamlining business processes too.

Figure 1: Example of a BPMN fieldwork workflow for a mixed-mode CAWI-CAPI survey design

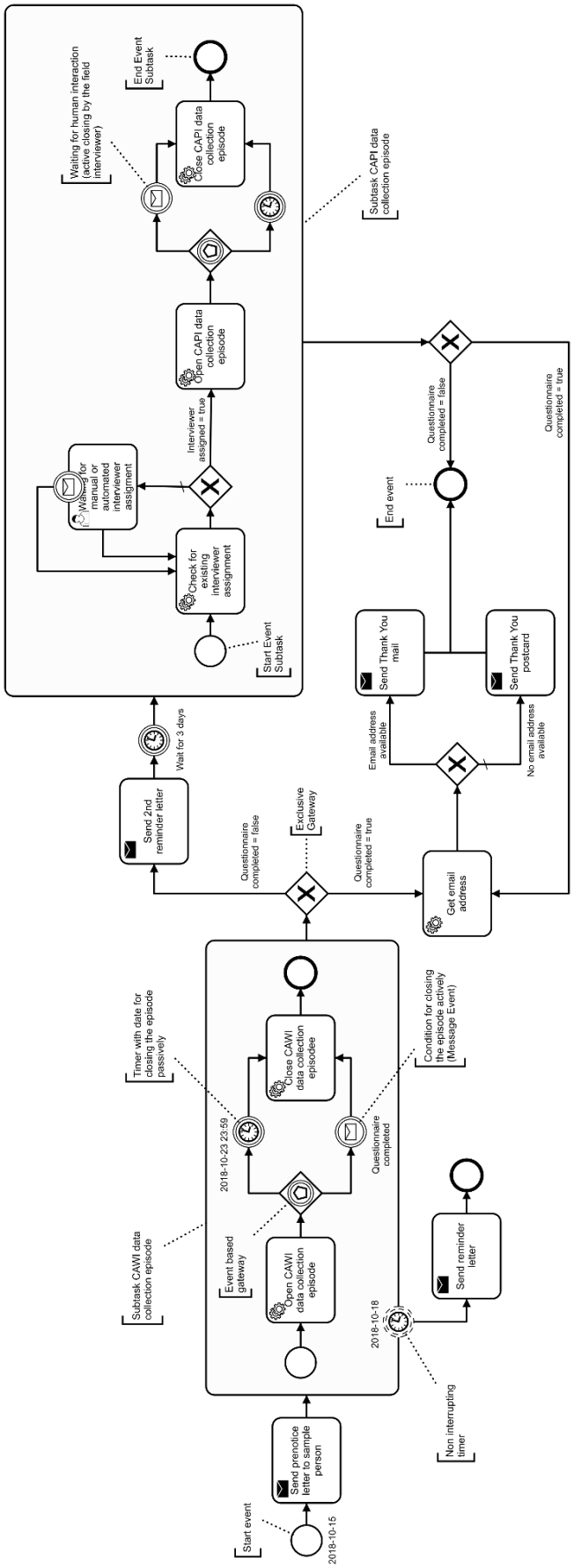
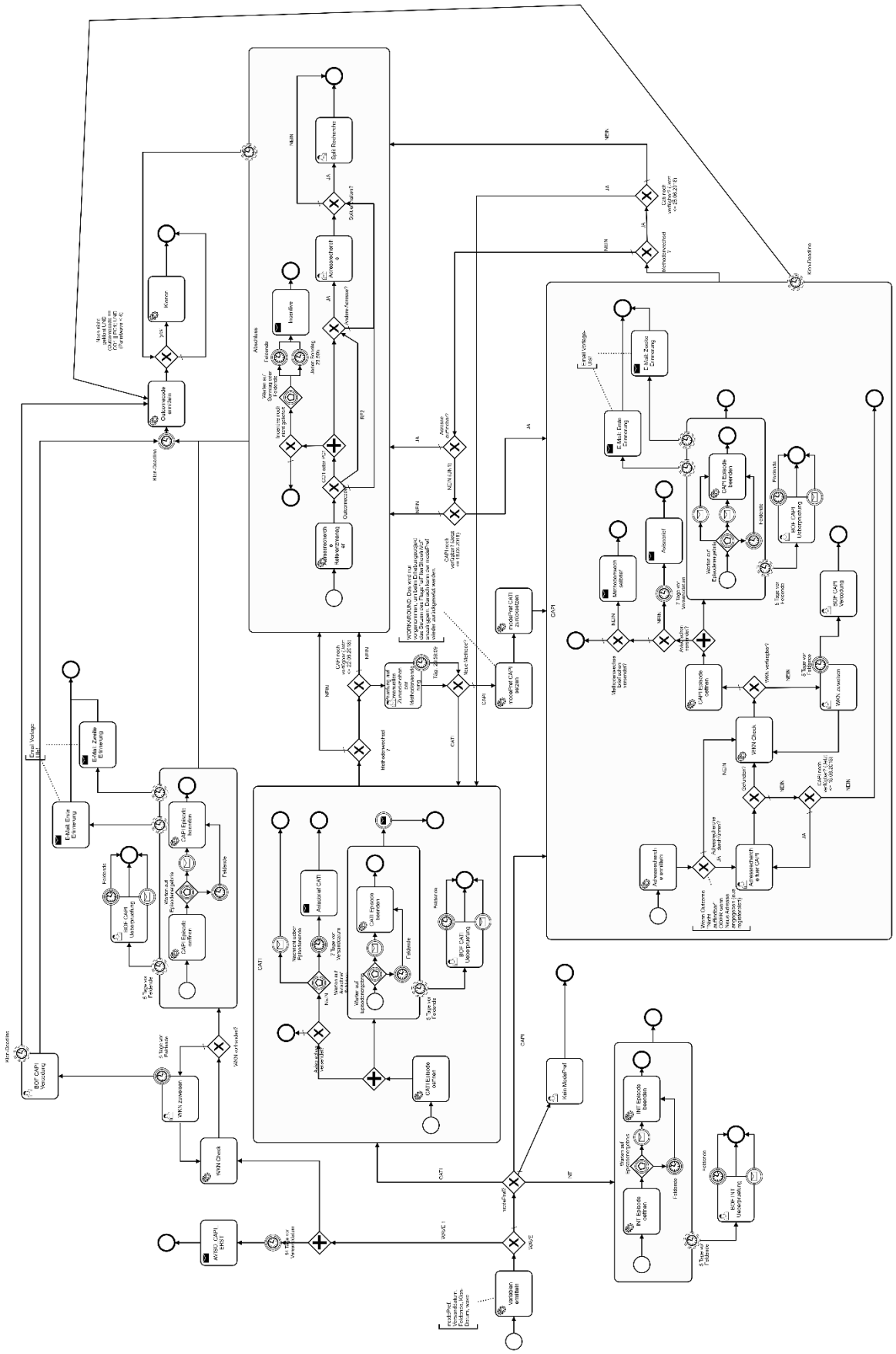


Figure 2: Default fieldwork workflow used for the Austrian EU SILC survey in 2018 running in the survey production tool STATsurv of Statistics Austria



The EU SILC survey (Statistics on Income an example and Living Conditions) has a rather complicated panel design resulting in complex business processes. The SILC survey is therefore a good example to demonstrate the complexity of a survey fieldwork workflow. Figure 2 shows the default fieldwork workflow used for the Austrian SILC survey in 2018 with a CAPI/CATI mixed-mode design, including a task to move a sample unit automatically to the sample of the next panel wave.

Using the Camunda modeler results in both, a graph for a survey fieldwork workflow and a fully functional XML-specification of that workflow (including the graph). This XML-specification needs to be uploaded to an application. To achieve an operational workflow the Camunda BPMN Process Engine has to be part of the service layer of this application handling all business logic and data storage. The IT service tool STATsurv supporting the fieldwork of social surveys and developed by Statistics Austria over the last years (see last section of the paper) meets these technical requirements.

Simplifying the fieldwork process to a sequence of mode-specific Data Collection Episodes and some pre-defined Communication Tasks implies that most of the human activities potentially occurring during the fieldwork phase are *not* controlled by an automated workflow. In case of a CAPI Data Collection Episode the interviewer does not need a formalised workflow for his or her activities. It is sufficient to provide an application with all information and tools needed to support and guide the interviewer in his attempts to contact the selected respondent and to carry out the interview. In case of a CATI episode the implemented business rules of the CATI software and the telephone agent drive the process of contacting and interviewing the respondent. For a CAWI episode the respondent itself has full control over the process. In any case all persons involved in a specific Data Collection Episode for a specific sample unit (field interviewer or telephone agents, respondent and staff member in the organisation running the survey) need to have a clear picture about the start date and the pre-defined deadline for reaching a completed questionnaire.

Modifying and overruling the default fieldwork processes

As a rule of thumb about 80 per cent of all cases normally follow a defined standard process from start to end without any deviations. This rule should also be applicable to the fieldwork process of a survey. That means, about 80 per cent of the people involved in the run collection process of a survey (interviewer, respondents, to some extent also case management staff) behave straightforward in a way expected and about 20 per cent do not. Ignoring cases that could not or will not follow the implemented and automatically executed standard processes is no option. This would result in a loss of quality and will definitely violate the principle of ensuring a positive relationship between the statistical organisation and data providers.

Per se, any automatisation of processes is cost efficient. Therefore, it is always favourable to model the “optimal” pathway through the fieldwork process in the default workflow according to the design of the survey. In addition, a workflow should also cover some pathways not desired but highly expected to occur. This strategy could reduce the proportion of non-automated case management (all case management activities for a single

sample unit not modelled already in the default fieldwork workflow) considerably. In practice, this strategy has its limits. Workflows get very complex and the risk of losing track of the processes in the workflow definition and in the actual fieldwork phase of the survey will increase. Once again, to control a certain task of the fieldwork process by an automatically executed workflow it is not always the best solution. Triggering a task by human interaction in a user interface of an application (a GUI) could be the much better solution.

To allow individual sample units to deviate from the automated default processes will anyway reduce cost efficiency of the survey operation. Therefore, one always has to find an acceptable compromise between some individual, non-automated case treatment and the associated cost and burden for the organisation running a survey. In an ideal world some business rules and a set of use-cases are defined already before the start of the survey and guide case management staff during the fieldwork phase. Finally, applying automated and non-automated case management should be feasible within one IT service tool to have a comprehensive and standardised documentation of all case management activities for each sample unit.

In the paragraphs following some most frequent needs for modifying or overruling certain aspects of a predefined default fieldwork workflow are presented and it is outlined how these cases are implemented in the service tool STATsurv. Again, not all requirements for non-automated case management need to be solved by a manual manipulation of the fieldwork workflow. For example, manually triggering the sending of a prenotice letter or a reminder letter to a sample unit once again (the first sending being triggered by the default workflow) should be a feature of the CAPI and the CATI agent applications as well as of the case management application.

According to the business concept presented above each Data Collection Episode addressing a single sample unit has a pre-defined length and therefore implicitly a pre-defined end date. For various reasons the interviewer or the respondent sometimes need additional time to complete a questionnaire. The staff responsible for case management could be informed about that need for additional time to complete the questionnaire in two different situations:

1. The Data Collection Episode for the sample unit is still ongoing.
2. The Data Collection Episode for the sample unit mode was already terminated.

In the first situation, a user of the Survey and Case Management Application of STATsurv searches for the sample unit, selects the default workflow graph for that sample unit, selects the timer symbol of the ongoing Data Collection Episode and modifies the predefined end date directly in the GUI according to the needs.

In the second scenario a more complex solution is required. It is not applicable to just modify a timer in a GUI. After termination of the Data Collection Episode the token (in the BPMN terminology the token is what moves through a process; for a survey workflow the token is a sample unit) continues to follow the designed workflow path. Depending on the respective fieldwork workflow, the token (the sample unit) could have entered a next Data Collection

Episode (with a different mode) or already reached the end event of the workflow. Simply moving back the sample unit to a previous Data Collection Episode is not possible in a sequential workflow.

To find a solution for that situation it is necessary to implement three features in the Case Management Application:

1. First, it must be possible to upload supplementary workflows in addition to the default workflow of a survey. These supplementary workflows normally contain specific sections of the default workflow.
2. A second feature necessary in the Case Management Application is the possibility to stop the actual running workflow for a selected sample unit.
3. A third feature allows for starting one of the (supplementary) workflows available for the specific survey for a single sample unit not actually being in a workflow.

To apply these features for a selected sample unit in a proper way demands a well-trained case management staff. All manual interventions have to be fully documented and commented in the paradata of the sample unit. However, manually stopping and starting different pre-defined workflows for a single sample unit in the Case Management Application creates high flexibility for case management. Moreover, an individually treated sample unit could be returned to an automated case management process after the intervention. For example, in the fictitious mixed-mode CAWI-CAPI survey mentioned above, the respective field interviewer of a respondent who is already in the CAPI Data Collection Episode informed case management staff that after all the respondent prefers CAWI for answering the questionnaire. Otherwise, he or she will not participate in the survey. Stopping the default workflow and bringing the sample unit to a supplementary workflow will satisfy this respondent request. The supplementary workflow needed consists of the first part of the default fieldwork workflow of the survey (the CAWI Data Collection Episode) without the request for a prenotice letter and ends after the thank you email or postcard.

Technically, it is possible to upload dozens of supplementary fieldwork workflows for a survey in advance or even during the fieldwork phase. For practical reasons the number of such supplementary workflows should be limited to a maximum of three to four. In an ideal situation, the field management staff always could decide in advance what needs for non-automated individual case management during the fieldwork should be satisfied with respect to quality and cost efficiency. In practice, for complex surveys with long fieldwork phases, it is not always possible to foresee all upcoming needs for intervention. However, the possibility of uploading additional supplementary workflows even during the fieldwork phase allows to adapt to upcoming needs and to do some reasonable adjustments in the use cases for non-automated case management.

Another frequent scenario is the need for a (final) outcome coding for a selected sample unit by case management staff. To be consistent, for this kind of outcome coding the concept of a Data Collection Episode is applied too. Such a special episode could be started in the Case Management Application manually if the selected sample unit is not currently involved in an open Data Collection Episode (independent of being controlled by a workflow or not). If

appropriate, case management staff could also enter the questionnaire of the sample unit in a selectable mode (as some parts of the questionnaire could be different for different modes) during that special episode. Case management staff has to terminate the episode indicating an accurate outcome category in the same way as field interviewer or telephone agents should do in their the web applications . All these activities are protocolled (which user is doing what and when) in the process data (paradata) of the sample unit in the same way as for all other mode-specific Data Collection Episodes.

The BPMN survey fieldwork workflow as a component of STATsurv

In 2004, Statistics Austria set up a survey infrastructure unit with a staff of about 10 to 12 people within the Directorate Social Statistics. In the first years, the unit was responsible for the fieldwork of the obligatory Austrian Microcensus (the Labour Force Survey) only. In the following years, more and more non-business surveys were carried out using mainly CAPI and/or CAPI modes. The unit serves a staff of about 160 self-employed field interviewer all over Austria and an outbound and inbound telephone studio with about 50 CATI stations. For the CATI agent and supervisor clients as well as for the questionnaires Blaise was used. For survey and case management as well as for all correspondence with sample units and with field interviewer a mixture of different tools (mainly MS Office applications) were implemented in addition.

This IT infrastructure tends to become insufficient as the number and the complexity of surveys increase over time. Moreover, using Blaise for CAWI was no option because of IT security reasons. In 2013, Statistics Austria started developing a new IT service infrastructure (STATsurv). Because business surveys use a well-established electronic questionnaire tool for many years in Statistics Austria, the new tool is focussing on surveys addressing persons living in private households. From the beginning of 2018, all social surveys including the Labour Force Survey and EU SILC are successfully carried out using the new service tool. The development process of STATsurv started from the scratch applying agile software development with a team of about ten to twelve people, among them five to six software developer. The development process was organised according to Scrum guidelines with equally shared responsibility between the Directorate Social Statistics and the IT-Division of Statistics Austria.

The general set-up of the new software focus on the following targets and standards:

- Use web-based applications for all user interactions.
- Embed all applications in the web portal of Statistics Austria for authentication and user role management of all internal and external user.
- Store all data in a central database.
- Involve as less IT staff as possible in launching and running a survey.
- Support automated and non-automated case management in an efficient way to allow for running several surveys with complex contact strategies and mixed-mode designs simultaneously.
- Promote the reuse and standardisation of questions by establishing a hierarchically structured repository (theme – module – submodule – dimension – question).

- Design questionnaires in a browser user interface.
- Use R scripts for all features not necessarily need a GUI.

Figure 3 depicts the basic architecture of STATsurv as fully implemented with the beginning of 2018. Ongoing developments are the integration of innovative smart phone survey applications (the first one being used for the Austrian Household Budget Survey starting in mid-2019) and the enhancement of the questionnaire component to be ready for a mobile-first approach for all questionnaires from 2021 onwards.

Figure 3: Basic architecture of STATsurv, an IT-tool for running surveys at Statistics Austria