Data Collection Management System in Statistics Finland

Pyry Keinonen, Joonas Salmi, Heikki Leino and Petri Godenhjelm Statistics Finland

1. Abstract

This paper discusses how Data Collection Management System was developed in Statistics Finland and in which way it is used in practice. Data Collection Management System was developed in 2017 to 2019 and was implemented to production in January 2019. System uses Blaise5 as primary Data Collection System.

Data Collection Management System in Statistics Finland consists of the primary data collection management service and from two smaller sub-services. The main service provides tools for survey, sample and case management and monitoring of data collection. One of the sub-services provides the user interface for case and interview management for interviewers, and one provides the infrastructure and API for web-survey data collection. Together these services provide the necessary functions for a multi-mode data collection process.

Statistics Finland has had a need to develop a data collection management system because the old operating environment has a lot of manual work stages that require effort. These workflows are largely built around the operating models of the Blaise4 production environment. The development of the Data Collection Management System has enabled production of reports on data collection and comprehensive management of multi-mode data collection. Most importantly, it enables online and offline interviewer data collection at the same time as online data collection.

The Data Collection Management System is built to utilize Blaise5 as a data collection tool for online and interview data collection. Therefore, following the Blaise5 Evolution path is crucial because the effects are directly reflected in the definition and implementation of the system development needs.

Future development work may include implementation of the Blaise5 CATI system and enterprise data collection features. In addition, the development of utilizing geographic data and information into the system to allocate cases to interviewers and reduce logistic costs is underway. In the future, the data collection management system will be used for all personal data collection and the Blaise4 production environment is being driven down.

2. Evolution of multi-mode data collection in Statistics Finland

Statistics Finland's data collection system on household surveys has long been Blaise4, which is mainly used by interviewers for field (CAPI) and telephone (CATI) data collection. In addition, the organization has experience with the Blaise IS system in web (CAWI) data collection. The transition from Blaise4 to Blaise5 began in 2015. Since then, the web data collection has been progressively implemented and migrated to the Blaise5 environment. The same year organization started a set of projects to develop a multi-mode data collection system and infrastructure. These projects were completed in January 2019 when Data Collection Management System was successfully implemented in production.

2.1 Early phases of multi-mode data collection

The first steps towards multi-mode data collection in Statistics Finland were taken in 2015 when the survey “Use of information and communications technology by individuals” was created with Blaise5. The pilot data collection was conducted the same year and the first production data collection was conducted in 2016. In practice the web data collection was collected by using Blaise5 and all other data collection management were provided with the data collection processes built around Blaise4.

Alongside these first steps, Statistics Finland began developing a multi-mode data collection system and infrastructure in 2015 to fully migrate to Blaise5. The system was developed in 2015-2016 and the first pilot was conducted in 2016-2017 but had to be discontinued due to technical problems. The first data collection used in the pilot was “Labor force survey”.

Despite the encountered technical problems in the pilot the “Adult education survey” data collection was carried out by using the multi-mode system in 2017. After finishing the first data collection it was decided that the current development needed a lot of redefining and reconstruction based on the experiences gained and the implementation project was put to a halt. This meant that transition from Blaise4 to Blaise5 was delayed and multi-mode data collection could only be done by combining the existing Blaise4 production processes with standalone Blaise5 web data collection processes.

3. Data Collection Management System

After gaining a lot of experience from the previous multi-mode project, a new project was launched in 2017 under a code name Ruuti. The organization set four aims for the Ruuti system which were reliability, usability, unity and flexibility. These aims were based on the idea that the system should be functional and reliable but also easy to use in both managing the data collection and performing the interviews. One importance was unifying the working tools, processes and methods of data collection including automatization of manual work phases. Also, the possibility to combine multiple data collection modes and software was considered an important feature.

3.1 System architecture

The core part of Ruuti system is Data Collection Management Service called Mixeri. Its primary function is to handle the data collection entity and communicate with linked sub-services. Currently there is two sub-services or channels which are Web Data Collection Service and Field Data Collection Service. Other channels are possible to be added later such as CATI channel. From communication point of view Mixeri is a client and different channels are each one’s servers.

The integration of the Ruuti system into the statistical production process takes place exclusively through the Ruuti system. If necessary, Mixeri and all channels can be placed in different domain areas of the infrastructure.

The channels are independent and each one is dedicated to only one data collection mode. All channels have their own dedicated databases and are responsible for their own special features which Mixeri does not know about. These features are for example the Blaise environmental duplication in Web Data

Collection Service and Offline capabilities in Field Data Collection Service. Blaise integration is done only within the channels.

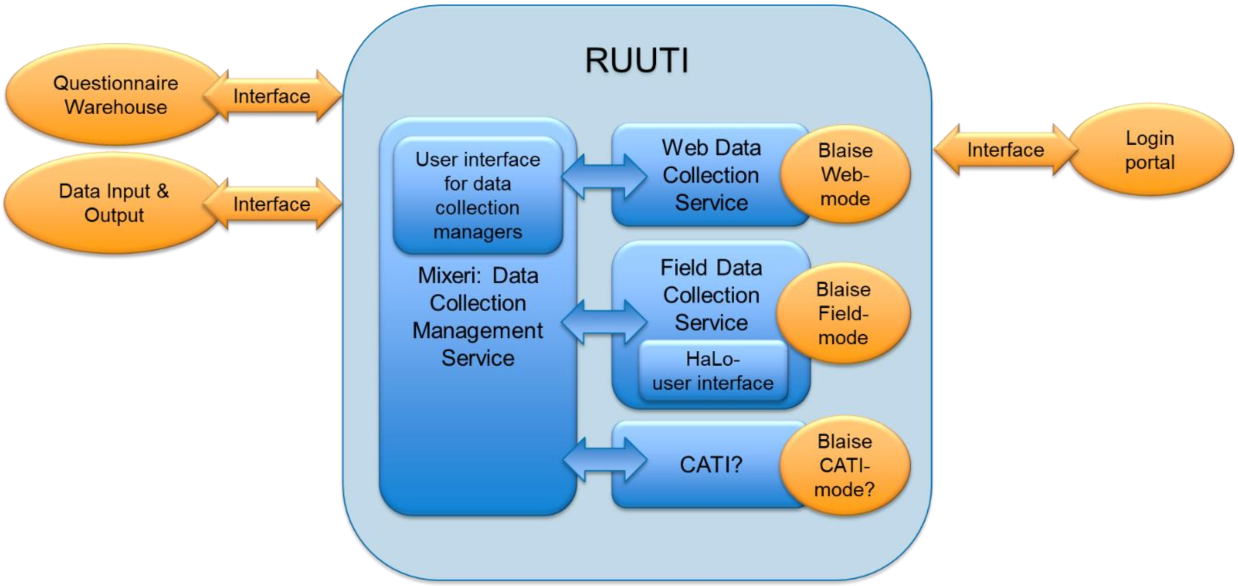

Figure 1: Ruuti Data Collection Management System

3.1.1 Mixeri – Data Collection Management Service

Mixeri is the central part of Ruuti System and it contains the knowledge of all data collections and cases and their status. Mixeri includes data collection establishment and management, case and sample data management including distribution, Blaise package distribution and management, interviewer resource management, and the management of data collection monitoring. It is a tool for data collection managers.

Mixeri communicates and synchronizes information necessary to channels. It enables the distribution of case data from all surveys to all interviewers and handles the multi-mode collected case data automatically. In practice this means that a case can be handled simultaneously in all data collection modes. In addition, a real-time overview of the progress of data collection and the ability to react to the situation during the collection is possible.

The most important user-side features are the possibility to create and monitor data collections and manage it even on individual data collection or individual interviewer resource level. Also, the distribution of cases to interviewers or from one interviewer to another is possible.

Together with Mixeri, Blaise, Web Data Collection Service and Field Survey Management Service provide the essential tools which makes a multi-mode data collection possible.

3.1.2 Blaise

Ruuti system utilizes Blaise5 as primary tool for data collection through Application Programming Interface. Blaise Server installation is used both in dedicated server for web data collection and individual interviewers' computer. Collected field survey data is stored in Blaise Database in interviewer's computer and web survey data is stored into SQL-server. In interviewer's PC the Field

Survey Management Service handles the collected field survey data and synchronizes the data to Mixeri when interviewer is online while Web Data Collection Service synchronizes the collected web survey data straight to the Mixeri's database. In both situations the initial process starts when Blaise Event triggers.

In case of the the survey package is changed in the middle of the ongoing data collection, Mixeri handles the possible data conflict situations. For example, if a new installed Blaise Survey package does not include a variable which is collected with the previously installed package then Mixeri keeps the previously collected variable data but does not pass it to Blaise any longer.

3.1.3 HaLo – Field Data Collection Service

HaLo is a Field Survey Management Service alias Channel and it provides the platform, communication between Mixeri and user interface for interviewers. It is installed locally to interviewer's workstation. HaLo completes the Blaise-integration by utilizing Blaise's API libraries. This integration includes starting the survey interview, case data pre-filling and reading the collected data after the survey session has ended. It also handles the installation of Blaise-packages and launches the Blaise5 survey.

The user interface of HaLo enables the survey and case management tools for interviewers. It works both online and offline state and synchronizes collected data and retrieves any updates between Mixeri and HaLo while interviewer has access to organization's network. The communication between HaLo application and Mixeri is handled by HaLo Server which is a service that passes messages in Field Data Collection channel.

3.1.4 Web Data Collection Service

Web Data Collection Service alias Channel consists of two different services which are WebHost and WebInstance. WebHost is a service which communicates betweet Mixeri and WebInstance and manages to which WebInstance commands are sent.

WebInstance completes the Blaise-integration in Web Data Collection channel by utilizing Blaise's API libraries. This integration includes the same capabilities as does HaLo in Field Data Collection channel. These are starting the survey interview, case data pre-filling and reading the collected data after the survey session has ended. It also handles the installation of Blaise-packages and launches the Blaise5 survey. All the necessary support services for identifying a case in login process is handled by WebInstance.

In Web Data Collection channel there is only one WebHost-installation that communicates with Mixeri. Channel may have more than one Blaise-installation. For each Blaise installation, there is also a WebInstance installation that communicates with WebHost. Currently there is only one Blaise-server-installation in use in Statistics Finland.

3.2 Interface services

Ruuti is utilizing other services through API for certain functionalities such as sample data input, collected survey data output and Login Portal for Web respondents. Also, Blaise survey packages are retrieved from external service. Some of these services such as Login Portal is used also in other Data Collection Systems in Statistics Finland.

3.2.1 Input & output

Samples for each individual data collection is uploaded to Ruuti System from statistical in-premise systems through API. One data collection may have many data collection periods and multiple samples. A sample always contains the ontology variables required by the system such as basic case information. Some of these variables are also passed to Blaise5 for example to be displayed in interviewer's survey view. Also, other survey-specific pre-fill variables that contain data is passed to Blaise5.
Answer data can be read from Ruuti Interface any time to be used in Statistical Processes. All the actual data processing is handled outside of Ruuti System. For example, the sample and pre-fill data must always be created in Statistical Processes outside of Ruuti System because the system itself produces only raw data.

3.2.2 Login Portal

Statistics Finland uses Suomi.fi e-Identification which is a shared identification service for public administration e-services. The service is in use in the national and municipal e-services, in which a user must identify themselves reliably.

After a successful login either using e-Identification service or using in-house created credentials the respondent is directed through Login Portal to the Web Data Collection Service. The service then launches Blaise5 survey with encrypted primary key in URL-parameter. The surveys actual primary key is decrypted before it is passed to Blaise. This process disables the possibility to share or hijack the case. For example, if the visible and encrypted URL is copied and reused, the decryption process catches this and blocks the entrance to the survey.

3.2.3 Questionnaire Warehouse

Questionnaire Warehouse is a service that enables Blaise Developers to upload Blaise Survey Packages into use. These packages are always built with unique Blaise GUID and named with unique Ruuti-ID. Also, the data model is named with the same Ruuti-ID. This way no problems will occur in case of the need to break Blaise data model and update the survey package to an ongoing data collection. When the Blaise packages are uploaded to Questionnaire Warehouse Service, they are instantly available for Data Collection Manager in Mixeri's user interface.

In case of the the survey package is changed in the middle of the ongoing data collection, Mixeri handles the possible data conflict situations. For example, if a new installed Blaise Survey package does not include a variable which is collected with the previously installed package then Mixeri keeps the previously collected variable data but does not pass it to Blaise any longer. If the Blaise data model changes dramatically then the new package should not be included to a data collection before the next period even though it is possible.

4. System implementation into production

Overall goal of going towards to digital data collection has taken a long step in last two years in Statistics Finland. This means different measures in survey communication starting from contacting respondents and reliable identification to the response process itself with questionnaires and ending to feedback and rewarding. There are four main areas where the change management was and is essential. The whole data collection process has been reorganized, the new data collection system is in production phase, and multi-mode questionnaire development and testing process is taking a new shape as Agile methods are being incorporated. Also, statistical methods are being developed further to answer the emerging analysis and quality issues.

In multi-mode administered surveys the demands of a flexible work division between interviewers is important and the tools for effective management and monitoring surveys have now been met with the Ruuti System. This development has been important at the transition from Blaise4-centered production to Blaise5-centered production. In our development process field interviewers and CATI interviewers use the same HaLo user interface in organizing their work. The feasibility of Blaise CATI management as part of Ruuti System is still being considered.

4.1 Roadmap of implementation

Ruuti System development started in August 2017. The first major steps were to rewrite a lot of the code written in the previous multi-mode system project and create a functional infrastructure and a functional communication between services. This meant largely the rewrite of Mixeri. The second step was to create HaLo Field Data Collection Service and its user interface. Third step included the creation of user interface and essential features for Mixeri.

In 2018 the first interview pilot survey was carried out and Ruuti System was tested. Later that year the work around Web Data Collection Service was started and it ended at the end of January 2019.

The implementation to production started in January 2019 but there were some compatibility issues with multi-mode surveys usage in the system. These problems were largely due to the inexperience integrating a demanding multi-mode Blaise Survey into the system.

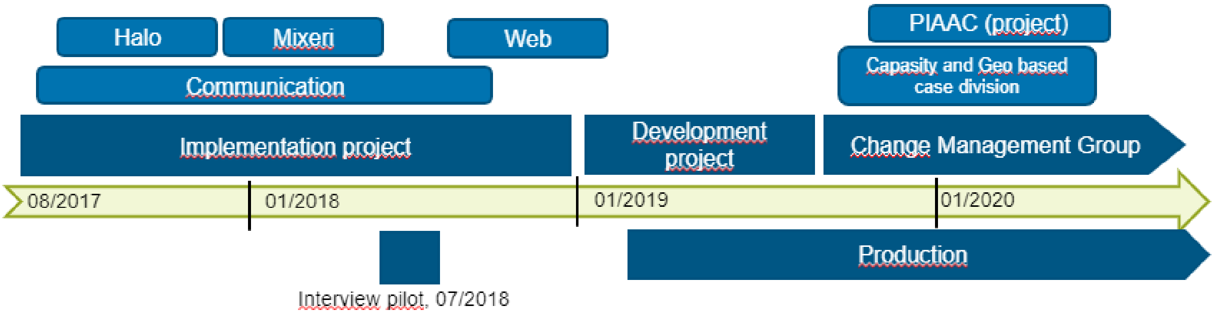

Figure 2: Ruuti System Development roadmap

4.2 Implementation schedule of multi-mode social surveys

Between the implementation years from 2017 to 2019, there have been three multi-mode pilot data collections carried out with Ruuti System. First pilot was carried out in June 2018 and its purpose was to test the HaLo Field Data Collection Service in action. The second pilot was carried out in November 2018 and in that pilot the technical capabilities for carrying out a multi-mode survey in Ruuti System was tested with Consumer survey.

Consumer survey together with Travel Survey were the first surveys to be implemented into production.

In 2019 another pilot was carried out when Labor Force Survey alias LFS was tested in Ruuti System. In this Pilot the formation of household with Blaise5 was tested the first time. In our roadmap for the coming year there is a major pilot for Survey of Income and Living Conditions alias SILC. In this pilot the household formation will be tested further.

Production wise the Time Use Survey will be carried out in August 2020. SILC and LFS are scheduled to be carried out in 2021 and Household Budget Survey in 2022.

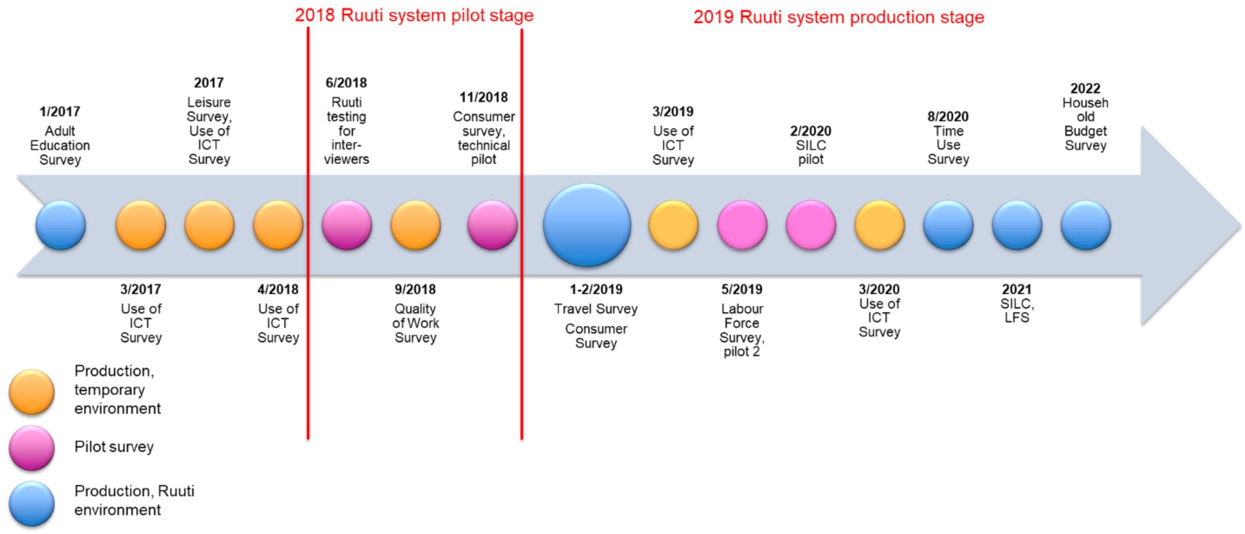
2018 Ruuti system pilot stage
2019 Ruuti system production stage
1/2017 Adult Education Survey
2017 Leisure Survey, Use of ICT Survey
6/2018 Ruuti testing for inter-viewers
11/2018 Consumer survey, technical pilot
3/2019 Use of ICT Survey
2/2020 SILC pilot
8/2020 Time Use Survey
2022 Househ old Budget Survey
3/2017 Use of ICT Survey
4/2018 Use of ICT Survey
9/2018 Quality of Work Survey
1-2/2019 Travel Survey Consumer Survey
5/2019 Labour Force Survey, pilot 2
3/2020 Use of ICT Survey
2021 SILC, LFS
Production, temporary environment
Pilot survey
Production, Ruuti environment

Figure 3: Implementation schedule of multi-mode social surveys 2017-2020

5. Conclusions and future plans

Development of the new data collection management was a much bigger effort than was considered at the start of the work. At same time moving from Blaise4 to Blaise5 created a lot of new requirements for the development process itself but also to the personnel of Statistics Finland with adaptation the new skills and processes. On business line of our data collection unit, the establishment of the dedicated product owner, was one big step to unify and intensify the more focused development process. The broader aims, reliability, usability, unity and flexibility to the Ruuti system are still valid. The time of evaluation and reflection will be relevant after we have moved all our production to Blaise5.

During system implementation some major flaws, bugs and technical issues of Ruuti system have been solved during 2019. In near future there are some new development areas in the roadmap of the system. Field Survey Management Service has got user feedback from interviewers. This will guide our development of more user-friendly user interface for interviewers. Other new important features of the system are the implementation of the Blaise5 CATI mainly because there is a special need to be able to use outside provider for interview services on the time of demand spike. And the possibility to establish and manage enterprise data collections is one perceived need.

We went to the implementation phase of system with MVP (minimum viable product) as defined by product owner. In future the work will continue to follow the principles of agile methods also in the questionnaire design. The new data collection system and Blaise5 together will guide the new generation of our statistical production.