GÉSA, the survey control system in Hungary
Frame, data collection, paradata and quality

Ildikó Györki[1]

[1]Hungarian Central Statistical Office, e-mail: ildiko.gyorki@ksh.hu

Abstract

Supporting the flow of survey design and data collection, the HCSO uses a standard metadata driven system, the so-called GÉSA system. This survey control system manages all economic and social statistical data collections of the office, observing the businesses and other institutions. The presentation introduces the central place of the system in the processing flow, the connection between registers, master frame and frames, the role of the system in the vertical and horizontal integration of different surveys. It highlights that the unified attributes and paradata make possible the standard monitoring, evaluation and quality assessment of the different surveys.

Keywords: survey, paradata, quality, frame

1 Introduction

In Hungary a general metadata driven system stands in the centre of the data collection systems from the middle of nineties. This so called GÉSA system manages all economic and social statistical data collections. The main principle of the system: the surveys, questionnaires are described in the metadata base, the survey frames have a common structure, during the data collection the same types of paradata are maintained with the same value sets (nomenclature), the functions are standardized, driven by metadata. This assures the unified monitoring and evaluation of the data collections and the effective development and operation of the system. The data collection, data editing and data processing functions are in close connection with the GÉSA system and survey frames. Because of the development of data collection instruments the importance of the different functions of GÉSA has changed during the almost twenty years, but the central role of the system and the unified approach of data collections remained steady.

2 The place of GÉSA in the processing flow

GÉSA (economic organisations and their data provision) is the survey control system for the observation of businesses and other institutions. It is the earliest metadata driven system in the HCSO which has been working since 1996. It is intended to give a general tool for supporting the tasks of the data collection phase of surveys, such as design, documentation, gathering of the questionnaires, and evaluation of data collections. The solution is built on standard procedures.

At present, GÉSA manages and controls 132 data collections which account for 98 percent of all data collections where the data suppliers are institutions. In the case of more than 80 surveys, the questionnaires

can be filled in and sent to HCSO via internet with proactive support. Besides data collections, 50 administrative sources belong to GÉSA for the sake of unified processing. The system maintains a population with almost 2 million units, more than 380 000 data suppliers and 2-3 million pieces of questionnaires in a given year.

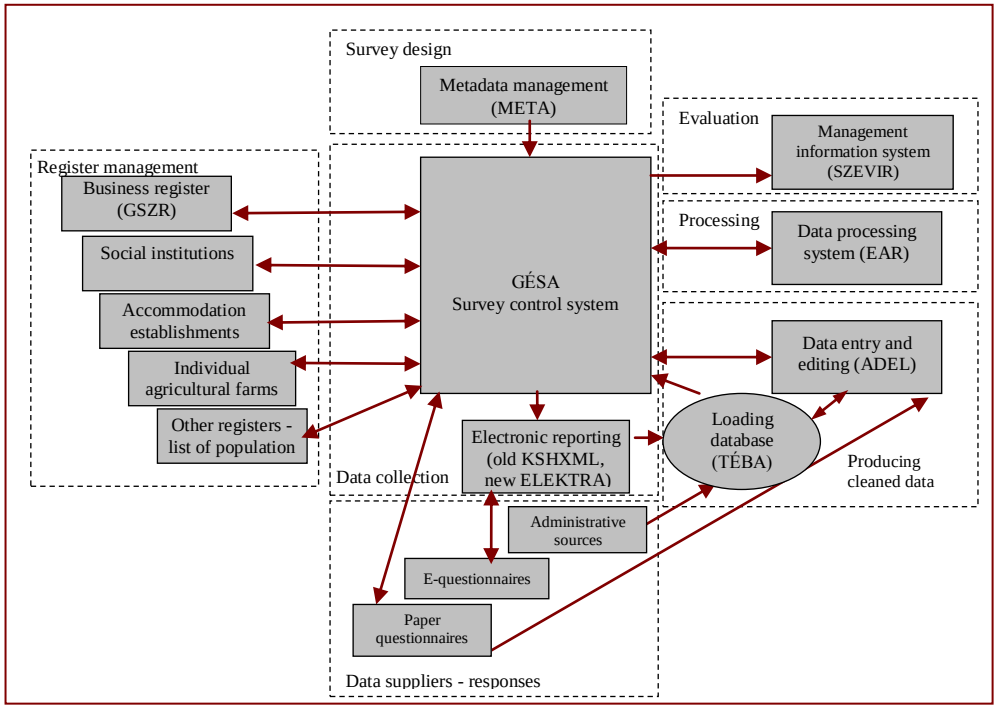
The GÉSA is built on the Business Register, its satellite and supplementary registers. In Hungary the Business Register covers the legal units with tax number (more than 1.7 million existing units), their local units and kind of activity units for the corporations over 50 employees. There are different other specialized satellite registers that are connected to the Business Register via the tax number like retail trade, social institution register, accommodation register, healthcare, research and development register, etc. One part of the satellite registers manages local kind of activity units. There are supplementary registers to the Business Register that contains not only the units with tax number but units without tax number like non-profit register with non-profit institutions and farm register with individual farms. These units without tax number are additional units to describe the whole population of economy. The GÉSA manages the surveys with population built on these registers. The interview type surveys built on the address register (population surveys) don't belong to the GÉSA, they are managed by the other survey control system (named LAKOS).

During the survey design the structural metadata of surveys have to be described every year: survey identification data, way and instrument of data collection, exact scope and forming rule of population, data supplier and statistical unit of the survey, sampling plan, way of mailing the questionnaires, detailed description of the questionnaires for personalization, deadlines of the phases of the data collection, etc.

The GÉSA functions are built on these metadata. The functions are performed for each reference period. The GÉSA supports the data collections via post, e-mail and web, the interview type data collection (for agriculture), and those secondary sources where the population is known in advance. The main principle of the functions is to give proactive support for the data suppliers to fulfil their duties. For the statistician colleagues whose task is the data collection the GÉSA gives support to deal with the subpopulations belonging to their responsibilities. The most important functions are the following:

- Forming the master frame from the snapshots of the registers (see the next chapter)
- Assigning the survey frames of the different surveys based on the master frame
- Selecting the sample of the representative data collections
- Selecting the questionnaires for mailing, their personalization, creating control information for printing, making a calendar for the data suppliers with their own response deadlines.
- Automatic and manual reminding and urging the data suppliers for response according to the urging plan by e-mail, fax and letter.
- Registering the responses, the way of responses, the non-responses and the negative answers and their cause.
- Maintaining and feed backing the changes and errors in the contact and other register information to the registers.
- Monitoring the flow of the data collection, to evaluate the result
- Computing quality indicators and response burden

Figure 1: The environment of the GÉSA system – relation among the statistical processing



GÉSA is in close connection with the different data collection techniques. For electronic reporting (primarily reporting on web now with the KSHXML, from the next year with the ELEKTRA system) the connection is direct, the GÉSA gives the duties of data suppliers, the frame information, statistical units, collecting units, personalisation and contact information for the respondents and for their e-questionnaires.

The questionnaires provided in an electronic way (web, e-mail, secondary source) are loaded into the data base. Their validation control and registration in the GÉSA is also unified and automatic (TÉBA system).

The paper questionnaires are entered to the data base by a frame system, ADEL. Its task is not only the data entry but it manages the data loaded from electronic source as well. The aim of this phase is to produce cleaned data. For the control, validation and editing of data the ADEL is in direct connection with the GÉSA.

The processing phase for imputation, estimation, aggregation and other analysis functions also use the survey frame information managed by the GÉSA.

In order to evaluate the surveys, to analyse the quality, to feedback the information to the survey design phase the GÉSA provides unified information for all surveys managed by it. It gives statistics and indicators about the data collections automatically and as frequently as it is demanded.

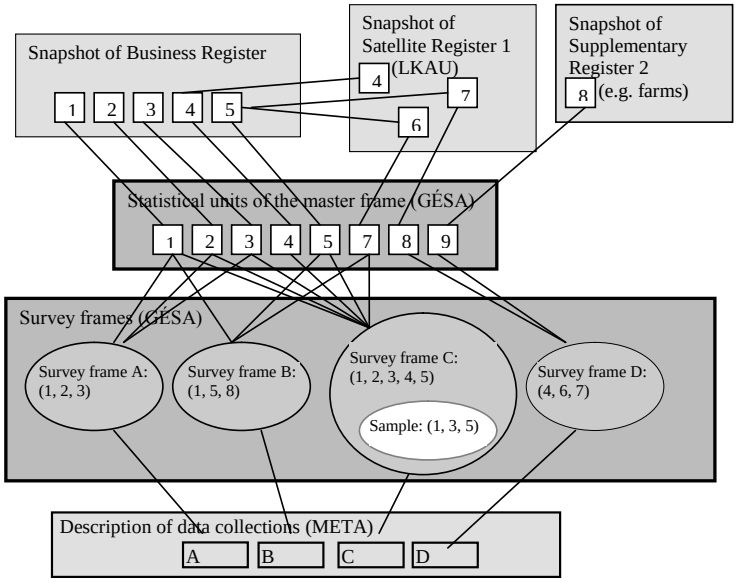
3 Survey frames and data integration

The frames of economic and social statistal surveys are created by the GÉSA system. The assigment of the data suppliers and statistical units is built on the *master frame* and metadata. The master frame is created according to the reference period of the surveys in every month and at the end of the year from the register snapshots referring to the given time. At forming the master frame the most important element is the Business Register, but the master frame contains not only the statistical units of the Business Register but the statistical units of its satellit and supplementary registers as well. The united statistical unites are labelled with unified identification number and statistical unit type.

During the survey design phase a precise description is made in the meta database about the scope and units

of the frame population (data suppliers, statistical units) and the connection with other surveys. The population is united from one or more subpopulations. The subpopulation can be selected by an algorithm on the attributes of the master frame unit. Beside that, other sources can be used to the assignment as well, like the result (data), experience (paradata) of the same or another survey from the previous period, or external sources.

Figure 2: Creation of the master frame and the survey frames



It is an important task of the survey design to prepare the integration of the survey data with the proper definition of the subpopulation of the surveys. We differentiate horizontal and vertical integration.

The first means linking statistical measures from different sources for a given population. For example, sales data of a retail trade data collection can be linked to the data of a labour survey. For the horizontal integration has to define common population (subsets of the population) for the surveys. These subsets are described and identified in the metadata base. The assignment of the population provides the same units for the subsets of the topics that we are planning to integrate. It is practical to select a common sample for the common subsets of the surveys because it helps the comparison not only for the estimated but the sample data as well.

In the case of vertical integration, we make a union of a given statistical measure for different, separately collected subsets of a particular population. For example, unifying data on the land usage of agricultural organizations, collected by self-enumeration with those of individual agricultural farms, collected by interviews creates data for the whole national economy. The base for the vertical integration of data is the definition of disjunctive populations of the surveys.

The selection of the survey frames from the master frame is automatic, built on the description of the population in the metadata base and the sources described. The assignment takes into account the relation of the surveys and the standard subsets of populations described above.

The survey frames and samples are stored and managed in one database table with common attributes for a reference period, so it gives an easy opportunity to analyse the survey frames, and the response burden. There is an application within the GÉSA system that makes data mining possible in order to analyse the different strata and units of the all populations together according to the response burden.

4 Unified paradata, monitoring and quality indicators

Information about the statistical units and data suppliers, the fault in their main attributes coming from the registers, the coverage error of the population, the important steps and the success of the data collection are collected and registered during the data collection on statistical unit level for all surveys and for all way of data collection methods. Where it is possible the information on the data collection phases is standardized, they are described by unified code lists (nomenclatures). In other cases textual information can be added to the data suppliers and statistical units about their readiness for response, and the changes in the unit that can be useful for the validation of the responses.

The most important attributes, paradata to monitor and evaluate the data collection phase of the surveys are the excepted and realized way of response (paper, e-mail, web), the steps of urging, the type of response (response with data and response without data), the reason of non-response or negative answer. The non-responses can be ranged into three groups:

- the first group deals with the frame problems: status of the statistical unit, data supplier (dead, under liquidation, etc.), classification problems, accessibility problems
- the second group deals with activity problems of the statistical unit, that has not the observed activity (now, ever, temporarily…)
- the third group deals with the respondents (it denies, is overdue, no successful contact, unknown cause)

During the data collection each questionnaire is characterized with these codes, if it is missing or it is sent without data. These codes are used at imputation and estimation to state who belongs to the survey population and who is not.

The information gathered in the data collection phase is used:

- for monitoring the progress of the data collection phase, to organise and control the work, to time the urging and other steps.

- for feed backing information about the register errors to the suit register.

- For determining the over coverage (unit doesn't belong to the population) and under coverage (unit must belong to the population, but formerly there was no information on it), and using in the next round of the data collection phase at the assignment of the survey frame.

- for evaluating the result and quality of the data collection.

There are statistics about the data collections automatically produced every day and every month. These inform about the rate of the different data collection instruments (paper, web, e-mail), the way of data entry (loaded data, manual data entry), the rate of the non-response. The detailed reports about the reasons of non-response, urging types, imputations, etc. are queried by the users of the GÉSA application.

The accuracy indicators of the data collection quality about the coverage, frame errors, the non-response, unit level imputation can be computed automatically.

References

Györki, I. (2012). GÉSA: The Tool for Survey Control, Quality Assessment and Data Integration. *Hungarian Statistical Review, Special number 15,* **48-78.**