KARAT: The new integrated data transmission system of the HCSO

Ildikó Györki, Ildikó Szűcs

Hungarian Central Statistical Office, Budapest, Hungary, ildiko.gyorki@ksh.hu
Hungarian Central Statistical Office, Budapest, Hungary, ildiko.szucs@ksh.hu

Abstract

The number of administrative data sources and other secondary data sources used for the production of official statistics has increased in the HCSO in the past few years. Data transmissions were managed differently throughout the HCSO with different security, metadata- and database-management. In order to address this issue, a new, integrated data transmission system – called KARAT – was developed in 2013-2014 for the transmission of secondary data to the HCSO from their data providers. The system has a number of automatic, proactive functions to support the tasks of the data masters and data providers of the secondary sources. KARAT also has monitoring functions to manage the whole data transfer procedure. The quality of the data transmission processes can be measured with different indicators. The new system contributes to the improvement, standardization and monitoring of the process quality related to the transmission of secondary data and data requests. These features make the KARAT system one of the key systems of the HCSO business processes.

Keywords: integrated metadata system, use of administrative datasets, secondary data sources, standardization

1. Environment of the KARAT system

1.1. The data collection system of the HCSO

Until quite recently, statistics in the HCSO was principally built on primary sources that is on data collections. Nearly 150 data collections provide proper data for the institutional (economic and social) statistics, and 16 data collections for population statistics. As regards the economic and social statistics, the main data collection mode is the electronic data collection via internet.

The statistical office uses a number of secondary sources as well, however, they have not been registered completely so far. The main part of the secondary sources are administrative (170 data sources), where the data registry has executive aims and legal bases. Another important part of secondary sources are statistical data collected by other organizations (85 data sources). Until now, the channels of transmission for datasets from secondary sources were quite different: e-mail, CD/DVD and download from public datasets (databases), etc.

Although the number and proportion of secondary sources are high in comparison to those of the primary sources, the support for the data collections (in documentation, standardization, application system development efforts) was higher than for secondary data sources since data collections demand higher expenditure (in terms of costs, organization and manpower).

1.2. The metainformation system of the HCSO

The description and documentation of the statistical system in the metadatabase have a long history in the HCSO. Data collections, their questionnaires and legal bases, the terms and concepts of the statistical system, the indicators in the statistical database, the nomenclatures, (classifications and other code lists) are registered in the metadatabase. The main characteristics of the statistical data processing steps such as data collections, data entry and validation, data processing, the building of data warehouse are also described in it. The above mentioned metadata are structured metadata, but for the methodological and quality description of the statistical domains descriptive metadata are applied as well. These are published on the HCSO website.

1.3. The HCSO's concept of application system development

From the nineties the HCSO's concept for the development of computer applications is that general systems – with standardized functions and driven by metadata – should be developed for the main steps of the statistical processing. Such systems were developed for data collection, for data entry and validation, for data warehousing and lastly for the data processing.

For example in the field of data collections there are two systems: one for the institutional data collections based on the business register, called GESA system, and one for the population statistics based on the address register, called LAKOS. These two systems organize and control the whole processing flow of data collections from the assignment of the survey frame, the sample selection, through the registration of the received questionnaires, the urging of missing questionnaires, to the monitoring and evaluation of the data processing flow.

1.4. The demand for and aim of the KARAT system

As we can see, the primary sources of the statistical system are managed by two general, standardized metadata driven systems, the GESA and the LAKOS system. However, for the secondary sources the HCSO had no solution. The secondary sources were not registered and documented. The channels of data transmission were not controlled and secured. The maintenance and archiving of secondary data was not assured, and the processing of secondary data did not always fit in the standard processing flow. Therefore, in 2013 the office decided to develop a general application system for managing the secondary sources and secondary data. This system is the KARAT system. The aim of the development was to create a system that

- carries out the documentation of the secondary sources;
- provides secure channels for the transmission of the datasets;
- supervises, controls and registers the takeover of the datasets;
- performs the formal control and check of the transmitted data;
- loads the arrived data into database for further processing and
- has proactive functions for supporting the data providers.

Besides the management of the secondary sources, there is an additional aim of the new system: to provide secure, documented channels for the data transmission towards external users that request data from the HCSO. This is a separate part of the system beyond the scope of this paper.

The application development of the system was finished in 2014 and it is operating from 2015 January. The involvement of the secondary sources into the KARAT system is on-going. Nowadays 183 of the 255 sources are managed by the KARAT system.

2. Metadata for the KARAT system

2.1. The data source register of the HCSO and the NSS

The KARAT system is built on the data source register. Before the introduction of the KARAT system, the aim of this register was to manage primarily the data collections of the National Statistical System (NSS). The register contained the information on the legal base of the data collections. This was the basis for the operation of GÉSA and LAKOS systems as well.

Documentation of secondary sources was not systematic and complete. In parallel with the KARAT development the data source register was redesigned to manage the secondary sources similarly to the data collections (primary sources). Now the register describes all statistical data sources of the National Statistical System in a unified system. It describes the data collections in the NSS, the administrative sources, other non-administrative sources, and the transmission of statistical data from other statistical organisations.

It identifies the sources and describes their characteristics such as enacting organisation, legal base, theme, population, frequency, size parameters, data collection mode, data master, data provider, privacy levels etc. It manages and registers the changes in the data sources.

2.2. Metadata for the secondary sources - KARAT metadata

The data source register is the base to describe the secondary data sources and to define the transmission tasks in the KARAT system. However, the register information is neither sufficient, nor is exact enough for the execution of the data transmissions. The detailed description of secondary sources are the so called KARAT metadata. The KARAT metadata give:

- the reference periods and deadlines;
- the type and structure of the accepted datasets;
- other attached datasets (documentation, code sets, etc.);
- special processing tasks related to the datasets (loading into database, completeness checking, etc.);
- the special attachments to inform the data providers (for example on the data structure);
- message types to be sent to data providers and statisticians responsible for the different steps of the data processing.

These metadata drive and control the data transmission flows and the functions of the KARAT system.

3. The KARAT system

3.1. The function of the KARAT system

The KARAT provides more secure channels for data transmission: user application for uploading the datasets on the HCSO website, and more system-system type interfaces (web service, IBM MQ WebSphere, governmental gateway). The main functions of the KARAT system are as follows:

3.1.1. Definition of the data transmission (secondary source)

Once a year the responsible statisticians create the plans for the takeover of the secondary sources. For new secondary sources they create, for repeated transmissions they check and, if necessary, modify the KARAT metadata (data providers, deadlines, accepted datasets, messages for the data providers and statisticians, processing mode of datasets, etc.).

3.1.2. Creation of the task items (organization)

Each month, in accordance with the data takeover plans, the items of the task lists are created automatically. An item serves to drive, control and register the events for a data transmission task for a given period of a given data source from a given data provider.

3.1.3. Data acceptance task

This function accepts the uploaded datasets, checks the completeness and the dataset formats, creates the tables for loading the data into database, checks the structure and loads the datasets into the tables, stores the accepted datasets for archiving.

The acceptance of a dataset is registered, and the data providers and the responsible statisticians are informed about the success or fault of transmission.

The accepted datasets either can be downloaded for the next steps of the statistical processing flow or accessed in the database tables by a general data validation and editing system, the so called ADAMES.

3.1.4. Message handling

In the KARAT system various types of automatic messages (emails) support the communication between data providers, statisticians and system administrators. There are

- informative messages about the obligations of data providers;
- reminder and urging messages for data providers;
- registry messages for the data providers and statisticians about the acceptance or the failure of uploading;
- messages for the administrators about system problems.

One part of the messages are connected to deadlines and others are driven by events of the system. Furthermore, there are standard letters prepared for the data providers, e.g. to initiate the resending of a new version of datasets.

3.1.5. Monitoring the data transmission flow

The system is provided with more functions to follow the status of data transmission tasks. It makes possible both to control a unique task and to create reports and statistics about the results and the quality of the selected transmission tasks.

3.2. The architecture of the KARAT system

In the KARAT system there are three applications for the users:

- The *external data reception application* outside the inner firewall for the data providers of secondary data. This application makes possible for the data providers to upload the datasets, to redefine the structure of their datasets if it has been changed and to update the contact information;
- The *external data sending application* outside the inner firewall for the data requestor of the HCSO for downloading the requested datasets;
- The *internal data reception and sending application* for the statisticians to manage the data transmissions. The users of this application can design and maintain their secondary sources, can follow and monitor the state of the transmission, download the accepted datasets if it is necessary, and administer the users, wrong/failed transmissions, etc.

From the three user applications the first and the third are related to the use of secondary sources which is the topic of this paper.

Beyond the three user applications, the background system functions – as the web service applications, the process control, the event and change management, deadline supervision, database loading, archiving, etc. – constitute a crucial part of KARAT.

4. KARAT paradata, information about the data transmission flow

During the whole process of data transmission from the design to the acceptance, the system collects and stores information in the database. This information implies the status and the dates of the different steps, events, such as:

- the status of the data transmission plan (pending or finalized);
- the date of creation and deadline for the data transmission tasks;
- the status of information sending to data providers about their obligations;
- the status of data transmission;
- the channel of transmission;
- the size and type of transmitted datasets;
- the acceptance status of datasets;
- the status of loading datasets into database;
- the degree of urging the data providers;
- the cause of missing data transmission.

This information is managed and stored in the database in a unified structure for all data transmission tasks, therefore it creates an adequate basis for the monitoring and evaluation. The above mentioned paradata make possible to create quality indicators and give information about the burden on data providers as well.

5. **Quality indicators in the KARAT system**

The quality of final output data is influenced by process characteristics, thus the improvement of processes is a key issue in the course of quality management. To ensure and enhance process quality, a regular and systematic monitoring system is needed. The key aspects that influence quality have to be determined, quality indicators have to be defined to control key elements, and regular, standardised measurements have to be carried out. To ensure quality, regular feedback on process quality has to be done and the reorganisation of processes has to be carried out when it is needed.

Besides the core functions the KARAT system has a function to monitor the quality of related processes. Two groups of quality indicators can be calculated in the KARAT system. The first group refers to quality indicators on secondary data sources, and the second to the measures of the different dimensions of data transmission processes. Quality indicators can be computed for the whole HCSO, by departments or by data owners.

The quality indicators of the first group show the documentation of secondary data sources and their involvement in the KARAT system on one hand, and give an overview on their complexity on the other.

5.1. Indicators on the documentation of secondary data sources:

- number of registered secondary data sources in the data source register;
- number and rate of secondary data sources involved in KARAT system.

5.2. Indicators on the complexity of data transmissions:

- number of secondary data sources loaded into database;
- number of transmitted files by data source and by period;
- transmitted data sets by file format.

Indicators in the second group refer to the quality of the data transmission processes. They measure different aspects of accuracy and timeliness related to the data transmission processes.

5.3. Indicators on the accuracy of data transmission process:

- rate of successfully realised data transmissions (number of successfully realised data takeovers divided by the number of planned data takeovers, multiplied by 100);
- number of erroneous data takeovers;
- rate of realised data transmissions (all data takeovers – sum of successfully realised data takeovers and of erroneous data takeovers – divided by the number of planned data takeovers, multiplied by 100);
- number and rate of missing data transmissions;
- number and rate of resent data transmissions (when the reason for resending is data error).

5.4. Indicators on the timeliness of data transmissions:

- number and rate of data takeovers until the deadlines;
- number and rate of data takeovers after the first urging message;
- number and rate of data takeovers after the second urging message.

It may happen that datasets are not transmitted to the HCSO even after the second urging message. The reasons for this – coded and stored by KARAT – can be diverse.

Quality indicators can be computed based on these codes. Regular analyses of indicators on missing data transmissions make possible to reorganise processes aiming to reduce the number of missing data transmissions.

6. Indicators on the burden on respondents

According to principle 9 of the European Statistics Code of Practice, European statistics have to be produced taking into consideration the non-excessive burden on respondents. To reduce burden on respondents statistical institutes use administrative data sources or other secondary datasets.

The use of secondary data sources reduces burden on respondents of data collections, however to some extent it increases the burden on the owners of datasets to be transmitted. Naturally, the overall burden on respondents for a transmitted dataset is much less than in the case of a primary data collection, nevertheless it cannot be assumed to be zero.

Based on the paradata stored in the KARAT system, indicators measuring the burden on respondents can be computed. Different data transmissions can generate different tasks for data owners, thus they cause different levels of burden. It depends on the number and complexity of datasets to be transmitted, on the channel of data transmission and on the amount and complexity of additional documentation to be transmitted with the datasets.

6.1. Indicators on the burden on respondents:

- number of datasets to be transmitted (by period and by year);
- channel of data transmission;
- documentation to be transmitted.

Based on the analysis of these indicators an optimal transmission structure can be defined for each data owner.

7. Conclusions

The introduction of the KARAT system ensures standard documentation of secondary data sources and data transmission processes. It applies uniform data transmission procedures and

provides a secure channel for receiving secondary datasets. The system also provides secure channels for the transfer of statistical data sets from the HCSO to users requesting statistical data. KARAT manages the loading of datasets to database and storing of all received datasets. It has a monitoring function to measure a number of process quality dimensions.

References

Manfred Ehling and Thomas Körner (eds) (2007): Handbook on Data Quality Assessment Methods and Tools, Eurostat. on-line accessed at: http://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf (27.04.2016.)