Integrated Metadata System in the Hungarian Central Statistical Office

Ms. Ágnes Almádi[1]
Mrs. Éva Hajósné Ender[2]

[1] *Hungarian Central Statistical Office, Budapest, Hungary; agnes.almadi@ksh.hu*
[2] *Hungarian Central Statistical Office, Budapest, Hungary; eva.ender@ksh.hu*

Abstract

In the Hungarian Central Statistical Office (HCSO) the concept of metadata driven, integrated statistical system has always been considered as a highlighted aim. Therefore the HCSO decided to elaborate a structured and integrated metadata system in the 1970s. The recent main improvement actions took place in 2007-2009. Owing to the development of the metainformation system currently all type of methodological data are provided by seven subsystems. Quality of metadata can be assured among others when they are up-to-date, comparable, consistent, available, standardized, related, complete, and understandable.

Ongoing standardization projects (like the Statistical Data and Metadata Exchange (SDMX) to standardize the structure of data-transfer; the Single Integrated Metadata Structure (SIMS) to describe the metadata of statistical domains; the Data Documentation Initiative (DDI) to describe the metadata of microdata; the Enterprise Architecture (CSPA)) confirm it to be necessary to overview the current metadata handling methods and metadata storing principles.

The aim of this overview is to improve the quality of the metadata and the efficiency in the integrated metainformation system of the HCSO. The classification of the statistical domains (the new version is based on SDMX List of statistical domains) is the base for the organisation of metadata. The purpose is to ensure the controllability of the maintenance through proper coordination, so the main aspects of the overview are the supervision of the

technical requirements, the content and the main observation factors of completeness and correctness of the metadata.

Keywords: integrated metadata system, standardisation, classification of statistical domains

1. Metadata driven, integrated statistical system – The past actions

The history of the development and improvement of the metainformation system of the HCSO goes back to the 1970s.

1.1. Establishing the foundation (1970s)

Baracza, Ercsey, Ábry (2009) mentioned that in the 1970s, the working groups of international organisations and the HCSO started focusing on research in the field of integration of information systems with metadata-related research included. The top management of the HCSO established a separate organisational unit for the improvement of the statistical information system so in 1974 the production database structure was finalised, to which metadata describing the statistical data content for users had to be connected. Metadata were also stored in MARK IV file management system in files. For the realisation of the system, an IBM mainframe and a batch process were available.

"The types of metadata connected to the database were the following: *1*. Hierarchy of statistical domains and files; *2*. Measures (including observed and aggregated variables) with their unit of measurement, periodicity, reference period and information on comparability; *3*. Nomenclatures (classifications); *4*. Nomenclature items; *5*. Variety of nomenclatures, which is the subset of elements; *6*. Nomenclatures determining the level of aggregation; *7*. Cross-references between nomenclature items; *8*. Statistical concepts." (Baracza, Ercsey, Ábry 2009, p. 107.)

Baracza, Ercsey, Ábry (2009) mentioned that in order to integrate other domains or fields into the system, their structure had to be planned because upload caused a problem, so the restructuring of the system was necessary which required great effort from statisticians and IT experts. The connection between the database and metadata was ensured by the naming convention for identifiers that time and the description of metadata was disseminated in catalogues.

1.2. The first metadata-driven system in the HCSO (1980s)

"In the beginning of the 1980s, a need arose for developing interactive accessibility of the data of the database for users so that they could make aggregation on data after choosing the right metadata. This demand was met by the statistical online data query system, called SOLAR (Györki–Papp [1985])" cited by Baracza, Ercsey, Ábry (2009, p. 108.). By this time the IBM mainframe became available from terminals so the system could be finalised after a long in-house development phase. When the users started to use it and gave positive feedback on its functionality, a lot of experience was gained on metadata needed to develop a metadata-driven system and also on the functionality required by a complex system.

1.3. The introduction of the database management system (1990s)

Baracza, Ercsey, Ábry (2009) summarised that in the beginning of the 1990s, there was an opportunity for the HCSO to renew its IT system with HP Unix operation system, ORACLE database management system and PC clients. The main task of this time period was the migration of data and the programming of systems. As the MARK IV files mentioned above and the files containing metadata were relational that time, they could be easily migrated to the new database, but new applications had to be developed for queries and maintenance works. Instead of migrating the above mentioned SOLAR system, the HCSO wanted to choose another system, which covered more existing software items but there was not enough capacity for the development.

1.4. Metadata-driven, integrated statistical system – The origin of the subsystems and the future development plans

Baracza, Ercsey, Ábry (2009) reviewed that the metadata system is a sub-system of the statistical information system and it covers the databases of metadata, as well as the activities and IT tools necessary for handling them. Metadata can fulfil several requirements, the most important one is to give information on the content and quality of data or on the methods of data production to the users; but there are metadata, which support, document the work of persons engaged in data processing. The automation and integration of statistical data production requires more and more parameters for operation which are also metadata, so the metadata-driven processes expand.

The HCSO currently has seven subsystems in the Integrated Metainformation System (IMS). According to the plans of the Methodology Department, the whole IMS will be reviewed in detail in the coming years to identify gaps and errors and to elaborate an adjusted development among the other subsystems. The aim is to ensure the completeness and the correctness of all new and existing metadata.

2.1. Statistical domains

Baracza, Ercsey, Ábry (2009) described that the terminology of several statistical domains was compiled in the 1970s and 1980s and these were disseminated on white papers. In 1995, the HCSO established a Methodological Working Group to elaborate the so-called basic documentation of statistical domains. As a result of the working group, the basic methodological ‘assets’ of the HCSO had been surveyed, and based on that the planning of statistical data collections became a HCSO presidential order in 1997. This work has laid down the foundation for the compilation of the methodological documentation of the subsequent statistical domains. “The ‘2005 HCSO’ strategy contained the task of supplementing the metainformation system with a new subsystem, which sets down the methodological background of statistical domains. Thus, supplementing the existing system with new metadata, it was established as an integrated part of the metainformation system. It is

available from September 2008 both in Hungarian and English on the website of the HCSO." (Baracza, Ercsey, Ábry 2009, p. 112)

Currently HCSO differentiates 126 statistical domains. Future plans for improvement: the HCSO is currently reconsidering the List of Statistical Domains from the point of view of the whole National Statistical System. This new structure will not contain the statistical registers (they are currently listed under the statistical domains and information on them is updated accordingly), and the processes. The description of the statistical domains will be upgraded according to the Single Integrated Metadata Structure 2.0 (SIMS 2.0) to support the Statistical Data and Metadata Exchange (SDMX) based data-transfer of the HCSO.

In connection with the improvement of the List of Statistical Domains the description of the processes will be structured according to the national adaptation of the Generic Statistical Business Process Model (Hungarian abbreviation: ESTFM).

2.2. Data sources

This subsystem was established in 2014, where all the metadata about administrative and non administrative data sources and data collections were merged into one unified database-table, called Data sources. In connection with the improvement of List of Statistical Domains, the description of the 'Data sources' will be also updated using the widespread international standard of the Data Documentation Initiative (DDI 2.5).

2.3. Legal base

This subsystem was established in 2008. The need for the improvement of the 'Legal base' arose from three sources: *1.* Since its establishment, no major improvements were carried out (it was high time for an in-depth review). *2.* The Hungarian statistical law is expected to change in 2016, which will clarify the legal bases of data sources and the data-handling through additional agreements. *3.* The development of the Integrated Data Request Management System (Hungarian abbreviation: ADKI) requires the legal bases to be added to each data request.

2.4. Concepts

Baracza, Ercsey, Ábry (2009) described that the need for completing the statistical concepts has a long history in the HCSO. One hand the IT background was already established for the maintenance of all concepts but on the other hand the enormous workload of uploading the approved, bilingual notions of all statistical domains were only carried out between 2007 and 2008. The work included the review of approximately two thousand concepts, which have definitions, cross-references, sources, and interlinks with other metadata. For further improvement of this subsystem, a 'supervision' was also performed by methodological supervisors about consistency and cross-references between the concepts of different statistical domains in 2009.

The need for the improvement of the 'Concepts' arose from three sources: *1.* Since the supervision and cross-referencing, no major improvements were carried out, *2.* a part of the existing concepts were not uploaded into the IMS. *3.* The Hungarian statistical law is expected to change in 2016, which will clarify key definitions and concepts for official statistics.

2.5. Nomenclatures and classifications

In the metadatabase of the HCSO, there are hundreds of nomenclatures both for all steps of the business processes. The last improvement took place in 2008. The need for the improvement of the 'Nomenclatures' arose from three sources: *1.* The principles for establishing the nomenclatures by the statisticians are loosely followed. *2.* Since the major improvement no overall revision was executed. *3.* A part of the nomenclatures are driving factors for systems such as the Integrated Data Processing System (Hungarian abbreviation: EAR), with their numbers increasing since the initiation of the Integrated Data Processing System (IT system) in 2010.

Baracza, Ercsey, Ábry (2009) elaborated that within the framework of the 2007–2009 general development phase, different classifications were selected for inclusion on the website. They were mainly international classifications (for example NACE, ISCO, COICOP, COFOG, etc.).

All the prioritised classifications were loaded into the IMS and a short description on these classifications (on their content, legal base, structure, history, applications, etc.) was prepared to give adequate information to users.

The last improvement is connected to the revision of NACE, CPA in 2008. A new database was created according to the updating procedure of classifications. The HCSO stores every year' updates on classifications in separate database tables so changes can be easy followed. The types of changes e.g. essential or syntactic are marked so the users can select them. In the explanatory notes of classification the Hungarian specialities are stored in separate cells of database table so the users can find their own activities in the classification.

2.6. Measures

The HCSO stores the basic metainformation of the measures, for example denomination, unit of measures, definition, reference period, frequency, level of aggregation. This subsystem could give a good opportunity for further integration, for example the operations and connections between varieties of measures could be very useful for the Integrated Data Processing System (Hungarian abbreviation: EAR) because it could help to define the level of aggregation. Nearly all integrated systems of the HCSO use the subsystem of the measures. Currently the measures are only visible in the Dissemination Database on the website. The stored measures are used during the whole statistical business process, and give the structure of the data as structural metadata.

'Measures' is the most idle subsystem of the IMS. The HCSO has plan to further develop this subsystem to store the operation between the measures.

2.7. Statistical registers

Currently 28 statistical registers are identified within the HCSO, but documentation on some of them is not publicly available. For the better understanding of data the HCSO publishes the documentation of many different statistical domains and statistical registers (8). It comprises the definitions of concepts, the data sources used in a statistical domain and statistical

registers, the methods applied, data quality aspects (user-oriented quality reporting), the most frequently used classifications and other metadata. The documentation covers statistical registers that describe the populations and the observed units which form the base for the statistical domains of the HCSO.

The basic information are the following: identification code, validity (from and till), description of population and order of the responsibility. Usually the statistical registers are stored in database. The statisticians and the IT experts mark the metainformation of recorded units and the connected attributes: name of the columns; role, type and validity of attributes, and the related classifications are written into the IMS. Also the connected statistical registers, number of the recorded units, validity of the relation and the related columns are identified. The frozen status is also saved on the last worksheet. The statisticians register the date of pre-frozen status and frequencies of pre-frozen.

The need for the improvement of the statistical registers arose from changing systems-needs, and users needs. In connection with the improvement of List of Statistical Domains, the description of the 'statistical registers' will be also updated using the widespread international standard of the Data Documentation Initiative (DDI 2.5). The Hungarian statistical law is expected to change in 2016, which will clarify the concept of 'statistical register' and will have provisions on the metadata of the statistical registers which has to be published.

2. Statistical production and metadata management

"Appropriate metadata management is an integral part of the statistical data production process. Metadata are present in each process phase whether they are generated in the process phase in question or recycled from the preceding process phase. Metadata management means the description of the statistical data and other outputs, e.g. methodological descriptions and documentation on the statistical domains, and provide parameters for the control of process phases and sub-processes. Metadata generation assumes special meanings in the individual phases because they are the building blocks of quality management in the sense that they provide a basis for the calculation of the individual quality indicators. The most important task

is that the necessary metadata are generated in the earliest process phase and be utilised in the later phases. From the perspective of the model the strategic issues of metadata management and their system are of vital importance." (HCSO, 2015, p. 166)

3. Quality Guidelines

In order to carry out HCSO's core activity in a quality manner, expectations should be clear to all, so quality guidelines are needed. It is clear to the participants that metadata are there in every process, so the HCSO describes the metadata management as an overarching activity for the business processes. The Quality Guidelines were updated in 2015 based on the previous version of the Quality Guidelines (2009) and the expertise accumulated by the HCSO since its publication. The following references were used to compose the new guidelines (HCSO, 2015): *1.* Eurostat's Quality Definition, *2.* LEG Quality recommendations adopted by EU Member States in 2001, *3.* the ESS Quality Assurance Framework adopted by the European Statistical System Committee in 2011.

The Quality Guidelines regarding metadata management are the following:

1. Metadata must be adjusted to user needs.
2. Metadata must be timely and topical.
3. Metadata must be available for users.
4. Metadata must be comparable across the various topics and with the various time series.
5. The consistency of metadata must be ensured.
6. When metadata are generated, international standards must be followed.
7. Standard metadata must be generated and used within the HCSO.
8. Metadata must be brought in line with the various topics and time series.
9. The completeness of metadata must be ensured.
10. Efforts must be made on ensuring integrity.
11. Metadata must be comprehensible.
12. If metadata use codes, an explanation must be provided for them.

13. Metadata must be documented.
14. Metadata must be identified and named separately.
15. Responsible persons must be allocated to metadata.
16. Metadata must be stable.
17. The necessary knowledge must be provided for the users of metadata.

4. Summary

HCSO has more than 40 years of history on the field of metadata integration where development cannot be finished, it can only be stopped at a certain point. Creative mind for implementing standards and general overview of the integrated system are necessary to conduct the methodological improvements.

The aim of the overview action to raise the quality of the metadata and the efficiency in a metadata driven, integrated system of HCSO. The classification of the statistical domains (the new version is based on SDMX List of statistical domains) is the base of the organisation of metadata. The purpose is to ensure the controllability of the maintenance through proper coordination, so the main aspects of the overview is the supervision of the technical requirements and the content and the main observation factors are the completeness and the correctness of the metadata.

5. References

HCSO (2015), Quality Guidelines - for the statistical processes of the Hungarian Central Statistical Office

[online available: http://www.ksh.hu/docs/bemutatkozas/eng/minosegi_iranyelvek_eng.pdf]

Baracza, G., Ercsey Zs., Ábry, Cs. (2009), Metainformation System of the Hungarian Central Statistical Office, Hungarian Statistical Review, Special Number 13,

[online available: http://www.ksh.hu/statszemle_archive/2009/2009_K13/2009_K13_103.pdf]