Data Warehouse for Agricultural Statistics as a tool for standardization of data processing and improving the quality of EAA calculations

D. Stankov[1], L.Tasevska[2]

[1] *State Statisical Office, Skopje, Macedonia; dejan.stankov@stat.gov.mk*
[2] *State Statistical Office, Skopje, Macedonia; ljupka.tasevska@stat.gov.mk*

Abstract

This paper provides information about re-designing of statistical processes in the area of agricultural statistics, aiming at increased efficiency and quality of the results and outputs produced in the State Statistical Office of Republic of Macedonia (SSORM).

The State Statistical Office of the Republic of Macedonia conducts various agricultural surveys and uses administrative data for statistical purposes. The activities are often decentralized and implemented in various ways, depending on the needs and the dynamic.

In this regard, and bearing in mind the quality issues, the SSORM has decided to establish a Data Warehouse for Agricultural Statistics. The data and metadata definitions are standardized so as to enable integration of historical survey data into a single database. The data model used for the Data Warehouse for Agricultural Statistics provides for the end user a simple and query-centric view of the data. The Data Warehouse enables an automatic calculation of the Economic Accounts for Agriculture data sets using available data from the surveys that are integrated in the Data Warehouse.

This paper discusses important issues related to establishing a Data Warehouse for Agricultural Statistics as a tool for standardization of data processing and improvement of quality of data, as well as increasing the efficiency in the working activities as a result of automation of processes related to compilation of data sets on Economic Accounts in Agriculture.

Keywords: Data Warehouse, standardization, metadata, data quality, automation

1. Introduction

The primary goal of the State Statistical Office of the Republic of Macedonia (SSORM) is to produce high quality, timely data harmonized with European Union (EU) requirements and to offer a solid basis for domestic and EU decision-making.

Stove-pipe oriented production, which implies “survey-dependent” solutions, is still being used in the SSORM. Every survey methodologist and every IT person is responsible for their survey and every survey has its own applications and databases. Taking into account the complexity of the overall statistical production and the fact that the SSORM has so far conducted around 300 statistical surveys, there are currently activities underway to change the system from a decentralized stove-pipe oriented production model to a more standardized statistical production model where resources are more efficiently used.

Improvement and standardization of the statistical production should contribute to reduction in the number of statistical systems and IT applications, efficiency improvements, reduction of risk, better documentation of the production processes, improvement of data dissemination, more flexible expertise and easier rotation of employees. In this regard, the SSORM has undertaken activities for development of the Data Warehouse for Agricultural Statistics.

2. Data model for Agricultural Data Warehouse (ADWH)

The agricultural sector is one of the most important sectors in the Macedonian economy, and therefore, agricultural statistics hold particular importance. Different data sets from different surveys and sources are produced within the SSORM, containing long series data from agricultural surveys and agricultural censuses, as well as administrative data. These data sets have varying and inconsistent data definitions or metadata across surveys. This absence of data integration and standardization is followed by an under-utilization of historical data, and impedes survey effectiveness and analysis.

In this regard and taking into consideration the importance of the quality of statistical data, the number of surveys and used sources, as well as the importance of the agricultural sector, the SSORM has developed the ADWH, an IT system that supports the processes of data storage, data preparation and editing, data processing, calculations, analysis, data comparisons, as well as preparation of the output tables and other formats for dissemination purposes. The calculation of the economic accounts for agriculture is the main output of the ADWH.

The ADWH integrates all survey data from agricultural statistics, survey responses and survey metadata. Metadata in this context is the information describing the structure of the questionnaires in a standardized and systematic template used for all statistical surveys. The data definitions, or metadata, are standardized so as to enable integration of historical survey data into a single database and management of metadata changes to surveys and census programs in the course of time.

In the data model shown below (Figure 1), facts are stored in the Survey Responses table of the model, which is the place of storage of the individual data response for a particular question from a particular survey or census. The Address List table contains information about survey respondents, i.e. it incorporates detailed information about individual farms or legal entities responding to a survey and/or census. In this regard, the ADWH data model enables link to the Statistical Farm Register, which serves as a data source for ADWH in terms of address data for the respondents and it is used for creating standardized and updated address lists that are unified for all surveys.

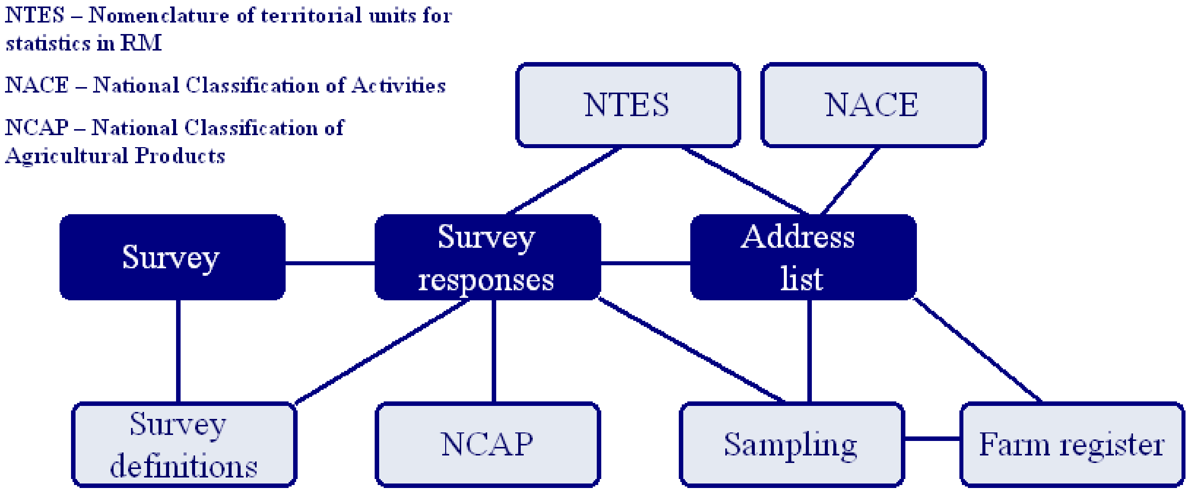
NTES – Nomenclature of territorial units for statistics in RM
NACE – National Classification of Activities
NCAP – National Classification of Agricultural Products
NTES
NACE
Survey
Survey responses
Address list
Survey definitions
NCAP
Sampling
Farm register

Figure 1: Data Model for ADWH

Survey definition tables are intersection tables that connect surveys and subjects participating in the surveys. Survey definitions, i.e., survey metadata are stored in tables that contain all information about surveys, such as survey name, code, reference data period, frequency, content of the tables and questions surveyed. A code from a current classification can be specified for each survey indicator. This enables achieving standardization, and gives the opportunity for analysis of survey results from different data sources and long time series linked by classification codes.

The official classifications and code lists used in surveys are stored and used in the ADWH. The system supports definition of new and changes of existing classifications and code lists. In order to provide better quality analysis over time, as well as creation of a strong system that fulfills all the goals defined, different versions of classifications are linked between each other, with corresponding tables. Likewise, there is the possibility to load additional classifications to new tables, and to use the same in the ADWH, as needed.

Additionally, to support the calculation of economic accounts for agriculture and price indices, there are satellite tables that are not directly connected to the main data tables of the ADWH, and most data contained in these tables are from various statistical and administrative sources. The system enables definition of measuring units for quantitative variables, predefined in the list of measuring units used in surveys and data collection.

3. Overview and outputs from the ADWH system

The ADWH supports data loading processes of the data-entry system from the statistical surveys and the electronic data collection system.

The data-entry system contains electronic data from various agricultural surveys for certain time series. The ADWH enables data loading and storage of the historical data available in electronic form. It also enables regular loading of data from new surveys. As surveys change over time, the system supports the change and enables definition and data loading of new surveys.

Data validation can be performed at a survey level (comparison between indicators and comparisons over time) or between several surveys as part of the data loading process and/or after the data aggregation process.

Data availability at micro level, to be used for statistical purposes only, enables transformation of the data to ADWH tables, rendering them appropriate for statistical analyses and reporting.

The system enables data transformations and data tabulations into tables best suited for further analyses and reporting. The data transformation is performed according to the transformation and aggregation rules also defined in the system. Transformation and aggregation rules can be established to reflect changes in input data, surveys or output requirements.

Finally, one of the benefits of high importance for the SSORM refers to the fact that the methodologists can prepare different statistical analyses defined according to their needs. With no IT support, by means of user-friendly statistical tools, users can perform various comparisons, cross-analyses, data-exploration, aggregations, interpretations and other statistical analyses.

In addition to data processing and data analysis, the ADWH also supports the process of data updating in the Farm Register based on new, more accurate data collected in surveys and processed in the ADWH system.

The system supports data dissemination, data visualization and report writing. The automation of the process of data dissemination is another advantage of the ADWH. Data tables in the ADWH are stored as datasets that are quite flexible and provide the ability to generate different types of reports in various formats available in the software.

Methodologists can create reports in standard office formats and produce output in a variety of markup languages. Reports for publications are defined with row and column headers and are bilingual - Macedonian and English. The report definitions are outlined, maintained and stored in the system and can be reused whenever there is need for it.

The new system is developed as a centralized system, where all the data are stored in a corporate data warehouse. The new approach is intended for unification of what logically belongs together by using advanced IT tools to ensure the rationalization, standardization and integration of the statistical data production processes.

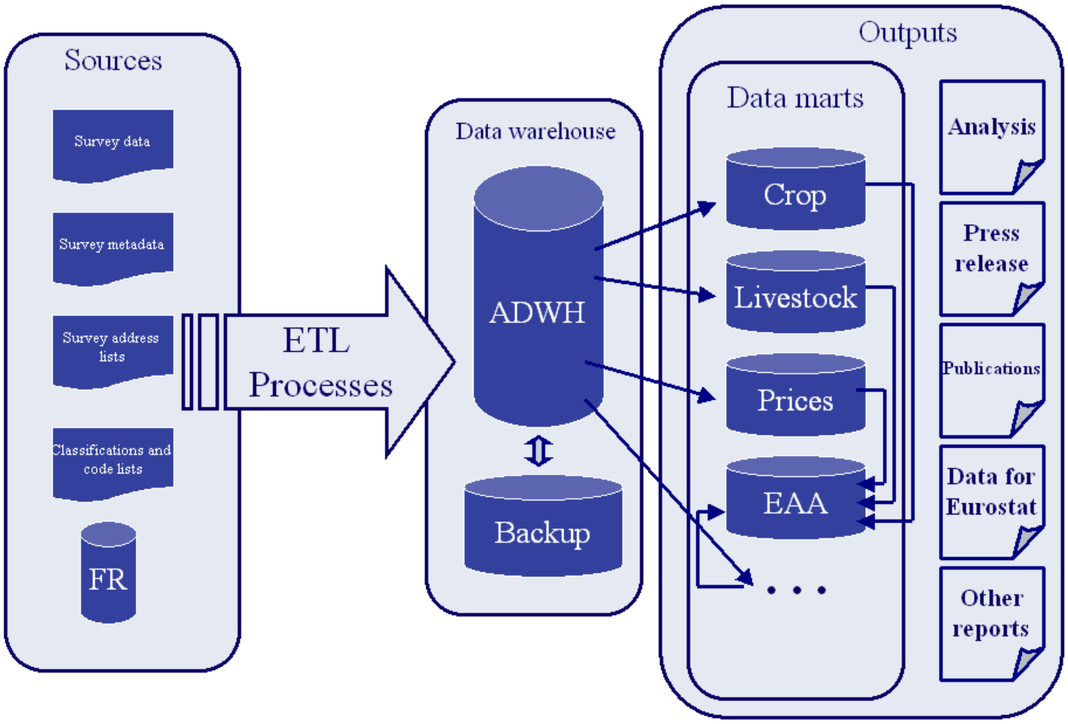

Figure 2: Overview of the agricultural data warehouse system

4. Calculation of the Economic Accounts for Agriculture into the ADWH

The ADWH is used as a tool for standardization of data processing and improvement of data quality. It enables greater efficiency in the work activities as a result of automation of processes related to data compilation. When we speak about efficiency, the process of compilation of data sets on Economic Accounts for Agriculture is significantly improved by using the positive aspects enabled by the Data Warehouse system. The process itself is automated, using integrated processes for calculation that are repeatable for each critical period of calculation. On the other hand, all survey data used as a data source for these compilations are already part of the ADWH and are available for the subject-matter staff for additional comparisons, crosschecking and analysis. In this regard, an important fact is that SSORM staff are trained not only to maintain the loading processes into the ADWH, but also to use, link, aggregate and appropriately analyze ADWH data with an aim to improve not only the quality of the basic survey data, but also the quality of the calculations made out of it, as it is the case with EAA calculations.

The process of EAA calculations is designed as a sequence of five interconnected steps / stages, each implemented as a separate project (Figure 3.). All stages are connected between each other, i.e., in each of them the outputs of their precedents are used as inputs and they have to be executed in the order shown in Figure 3.

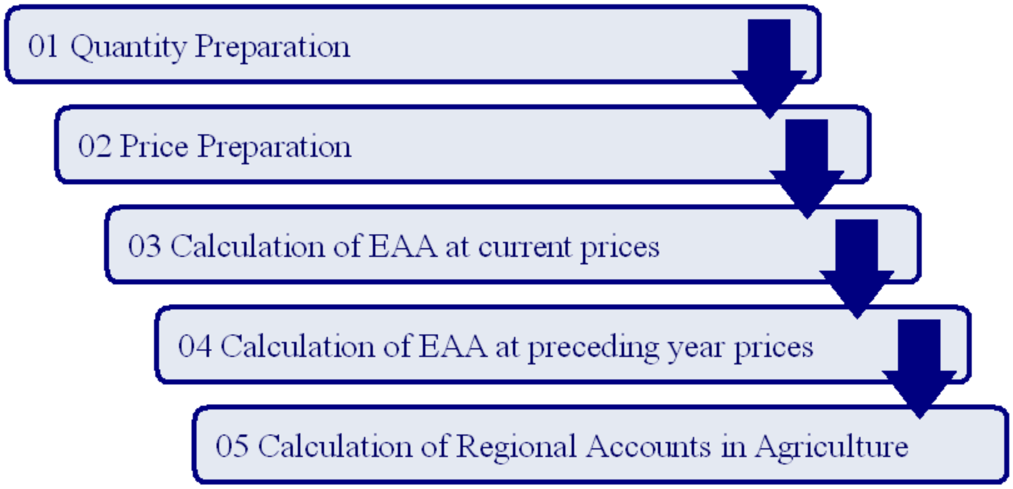
01 Quantity Preparation
02 Price Preparation
03 Calculation of EAA at current prices
04 Calculation of EAA at preceding year prices
05 Calculation of Regional Accounts in Agriculture

Figure 3. Process stages designed for EAA calculations

In the first two stages of the calculation, the processes related to data preparation are executed, while the last three stages are designed for the purpose of calculations of EAA outputs related to calculation of Economic Accounts for Agriculture at current prices, Economic Accounts for Agriculture at preceding year prices and Regional Accounts for Agriculture. The data on production are obtained from the statistical surveys conducted in the SSORM, as well as additional tables loaded in the ADWH for the purpose of these calculations. All of them are stored as data sets in appropriate libraries separately designed for each reference year. There is another library that contains mainly the tables that are not specific for a particular reference year and are not expected to be revised annually.

The *first step* "Quantity preparation" is based on the survey data regarding crop and livestock quantities at product level. Data from crop and livestock surveys are extracted and transformed into this phase to adapt their form to be usable for the purpose of EAA calculations. The data table prepared as result of the first step remains in the further process as an input for further calculations.

The *second step* "Price preparation" uses data from statistical surveys already loaded in the ADWH, as well as other data on prices prepared in a suitable format and loaded in the ADWH for the purpose of calculations. As a result of this phase, two tables are obtained, one for the calculation of EAA at current prices and one for the EAA at previous year prices report that will be prepared the following year.

The *third step* "Calculation of EAA at current prices" gives as a result the first data output named Economic Accounts for Agriculture at Current Prices. It is based on the quantity and price data produced in the first and second step, as well as data on other parameters, prepared and stored in the ADWH each year. These parameters are from statistical and administrative data sources. As a final result, the first report is produced at the fourth hierarchical level (data of the fourth hierarchical level are summarized in three steps), in order to obtain the final report, as it is required for dissemination purposes. This report may be changed upon further need of the Office.

A second EAA report is produced in the *fourth step* of the calculations "Calculation of EAA at preceding year prices". This report is based on the quantities data obtained in the first step, as well as prices data produced in the second step in the year t-1. As it was done in the third step, this report is also generated at the fourth hierarchical level, in order to obtain the required form of the final report named Economic Accounts for Agriculture at previous year prices. This report may be changed, if needed.

The final, *fifth step* results with the third report in the calculations, named "Calculation of Regional Accounts for Agriculture". For the purpose of this report, data tables from the second step (data on prices) and data tables from the third step (EAA at current prices) are used. This report is also produced, as it is required for its dissemination purpose. This report may be changed, if needed.

All of these steps are programmed and built into the system, where for the purpose of the calculations for the next year only some elements should be adjusted, i.e., input tables to be changed with the corresponding ones from the current year, all the processes to be executed and checks for possible errors in the system to be performed. As a final result, the three reports needed for the purpose of EAA calculations are generated in a more accurate manner, the data are with better quality, the time for calculation is significantly shortened and the disseminated statistical products are prepared in a standardized way. If needed (as a result of methodological changes, new data source, user requests, data revisions, etc.), the system is flexible and changes could be integrated very easily in the process.

5. Conclusion

The paper presents activities that have recently been carried out at the SSORM, aiming at development and implementation of the ADWH, as well as automation of the process of EAA calculation.

Several phases of the statistical production from the statistical business process model have been standardized by establishing the ADWH.

The availability of the databases to methodologists by means of user-friendly IT tools, with possibilities to perform deeper statistical analyses not depending on the IT staff, is one of the strongest features of this tool. Also, through the ADWH, the methodologists are able to share their work and analyses performed in the ADWH and may enable other colleagues to reuse the same analyses or to redo them on other data sets available in the system. The ADWH supports the process of data updating in the Farm Register based on new, more accurate data processed in the system. The improvement of data dissemination and data visualization is another strong side of the ADWH. The automatic creation of the dissemination outputs and Eurostat transmission outputs is performed in an automated and standardized way, and this also shortens the time for their preparation.

The re-use of data and the automatic calculation of the economic accounts for agriculture is the main output that incorporates almost all the surveys related to agriculture and imported in the ADWH. Furthermore, the punctuality, the accuracy and the quality of these calculations will also be improved, and at the same time, the data will be stored and archived on the server. This will enable the staff from the subject-matter department to be involved in the process of their calculation and validation.

6. References

ASA Institute, (2002) “User’s Manual for “EAA – OPAL”(Economic Accounts for Agriculture—Operative Policy Analysis) Version 1.2”

BercebalJ.M., Maldonado J.L., Martínez-Vidal M.A., and Salgado D., (2015) “Data collection and selective data editing in a systematized and integrated way: an experience in progress at Statistics Spain”; paper presented at the UNECE conference of European statisticians, 15-17 April 2015;

European commission, (2000) “Manual on the Economic accounts for agriculture and forestry rev.1.1”

Özturk A., Özyrek M., Tuncel.,Guven I., (2015) “Harzemli, the DDI based statistical production platform” paper presented at the UNECE conference of European statisticians, 15-17 April 2015;

Seljak R., Smukavec A., (2015) “Modernisation of statistical processing at SURS”; paper presented at the UNECE conference of European statisticians, 15-17 April 2015;

Speh T., (2015) “From data store to data services-developing scalable data architecture”; paper presented at the UNECE conference of European statisticians, 15-17 April 2015;

SSO, (2012) National Classification of Agricultural Products, 2011

SSO, (2013) NKD Rev.2, National Classification of Activities

SSO, (2013) NTES - Nomenclature of Territorial Units for Statistics

United Nations Economic Commission for Europe, (2013) “Generic Statistical Business Process Model”.

www.ec.europa.eu/eurostat