UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working paper
23 February 2015

Workshop on the Modernisation of Statistical Production
Meeting, 15-17 April 2015

Topic (ii): Enterprise Architecture and its role in the Modernisation of Statistical Production

MODERNISATION OF STATISTICAL PROCESSING AT SURS

Prepared by Rudi Seljak , Andreja Smukavec

rudi.seljak@gov.si, andreja.smukavec@gov.si

Statistical Office of the Republic of Slovenija

I. Introduction

1. Statistical data processing has always been a demanding, time consuming and consequently a very expensive task. This is especially true in the case of official statistics where due to possible political and administrative consequences the reliability of the statistical outputs is of special importance. It is therefore understandable that a lot of resources have to be spent on data processing, especially for the processes such as data validation, statistical data editing, aggregation and standard error estimation. Furthermore, confidentiality has become a very important issue in recent decades, introducing statistical disclosure control as another very demanding part of the statistical process. On the other hand, there is also a constant pressure for budget cuts, which is of course in evident contradiction with the above mentioned demands. The official statisticians are hence increasingly facing the challenge of producing confidential statistics of high (or at least sufficient) quality with the significantly reduced resources.

2. To at least partly reduce the gap between the above mentioned demands, in recent years a lot of effort has been put into the rationalization of the statistical process. One fact that certainly acts in favour of these efforts is the enormously rapid development in the IT area, meaning the development of hardware equipment as well as the development of a wide range of software tools, which are certainly at disposal to a larger and larger extent. So there is no surprise that also in the area of the official statistics in recent years a lot of effort has been made in the direction of efficient use of all these new tools and applications in order to make the whole production cycle less burdensome and most particularly less expensive.

3. At the Statistical Office of the Republic of Slovenia (hereinafter SURS) systematic work in this area began some seven years ago when the first prototype system for the modernized data processing was built. The prototype consisted of a few modules which aimed at “covering” the different parts of the statistical process (e.g. data validation, data correction and imputation, aggregation and standard error estimation, tabulation). We called these modules building blocks. From then on these building blocks were gradually developed and supplemented and have so far been successfully used in several large and demanding

surveys, such as the 2010 Agriculture Census, the 2011 Population Census and the EU-SILC survey. Although these applications worked quite well in the case of the above mentioned surveys and can by their main characteristics be denoted as the generalized tools, there are still a lot of features which characterize them as local solutions. To make a step further, SURS launched a project aiming at upgrading the existing solutions and building one global solution which would cover most parts of the data processing and which could easily be used for most of the statistical surveys.

4. The paper describes the gradual development of the generalised system for data processing at SURS. We begin with the first version of the new system, designed as a prototype, but already successfully used in some large and important surveys. Then we present the improvements of the system that are at the moment introduced through the implementation of the large infrastructure project called SOP . We conclude our story with the reflections about the necessary changes at the institutional level that will have to follow the modernisation of the statistical process and that will have to be introduced in the near future.

II. First version of the generalised system

5. Until a few years ago SURS used only stove-pipe oriented production, where solutions were “survey-dependent”. Every survey methodologist was responsible for his or her survey. How the statistical process (editing and imputations, sampling error estimation, statistical disclosure control, tabulation) was organized very much depended on the survey methodologist's experiences, organization skills and striving to improve the quality of products. One such example is statistical disclosure control for dissemination tables, where the methods for statistical disclosure control were applied only if the survey methodologist asked for tabular protection. The same situation was with the sample error estimation and editing; the procedures were applied only if the survey methodologist asked for the application to be developed. The survey-dependent approach was also (for most of the processes) used when the development of the software solutions was concerned. Software solutions were developed mostly ad-hoc for the needs of the particular survey. Such a system demanded a large amount of IT work at the development stage, making the maintenance of these software solutions a very demanding job.

6. Taking into account all these facts, we decided to build up a new strategy for modernization and standardization of our production system. The main goal was unification and standardisation of the statistical processing. There were two cornerstones of the strategy: development of the generalised metadata driven software solutions and creation of a database of process metadata. Generalised metadata driven programs that can be used in different surveys should be developed and process metadata for each survey will be saved in one place and transparently available to anyone interested.

7. To achieve this goal, we decided to break our statistical process into a set of smaller generic solutions, which should be designed in a way that they enable easy and flexible linking of inputs and outputs of the individual components to the whole statistical process. These components (the building blocks) should provide the generic software solutions for the certain parts of the statistical chain and should be designed in a way that they can act independently. The main features of these building blocks could be summarized as follows:

a) They are designed on the basis of harmonized, transparent and widely accepted methodological principles, which have been determined before the actual creation of the particular building block
b) They should be opened to such extent that these building blocks can be plugged to different databases in different environments (e.g. ORACLE, SAS) as long as the databases follow some basic rules for the organization of the data
c) They are designed as fully metadata driven systems, meaning that information which determines the parameters for the execution of the processing for a specific survey and a specific reference period should be provided outside the core computer code. No information referring to specific survey execution should be incorporated into the general program code but should be provided by the subject-matter personnel through the special metadata tables
d) The process metadata can also be provided in different databases for each survey in different environments, but each of these (metadata) databases must follow the strict rules of its structure (tables and variables)

8. In 2007-2010 the first version of the new generalised system was build. In the first version of the system for data processing, the input table consists of all the microdata to be processed, and this table has to be a SAS (work) table. The output of the building block is also a SAS table. Therefore, we always need a small ad-hoc program which transfers the output table back to the microdata database in case of data editing. A simplified schematic presentation of the functioning of a building block "data editing" is presented in the following figure:

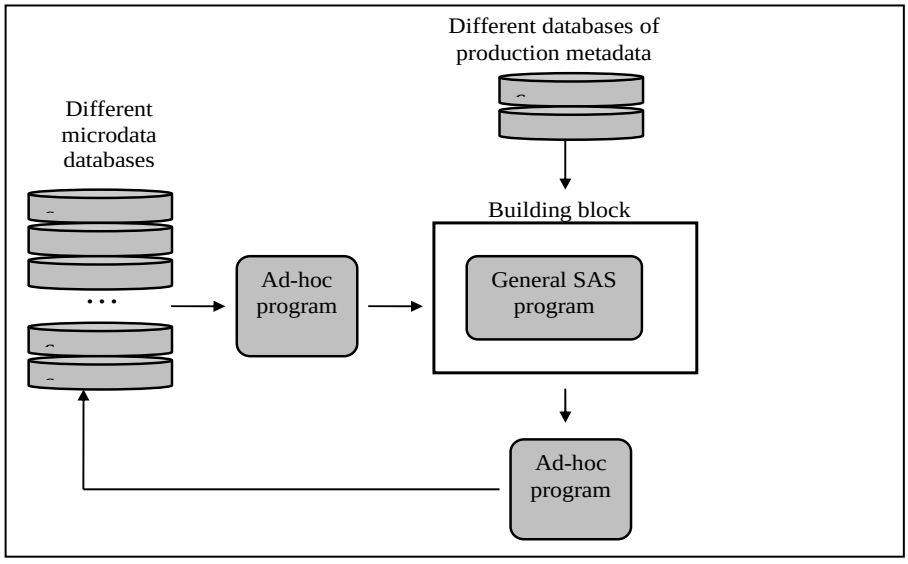
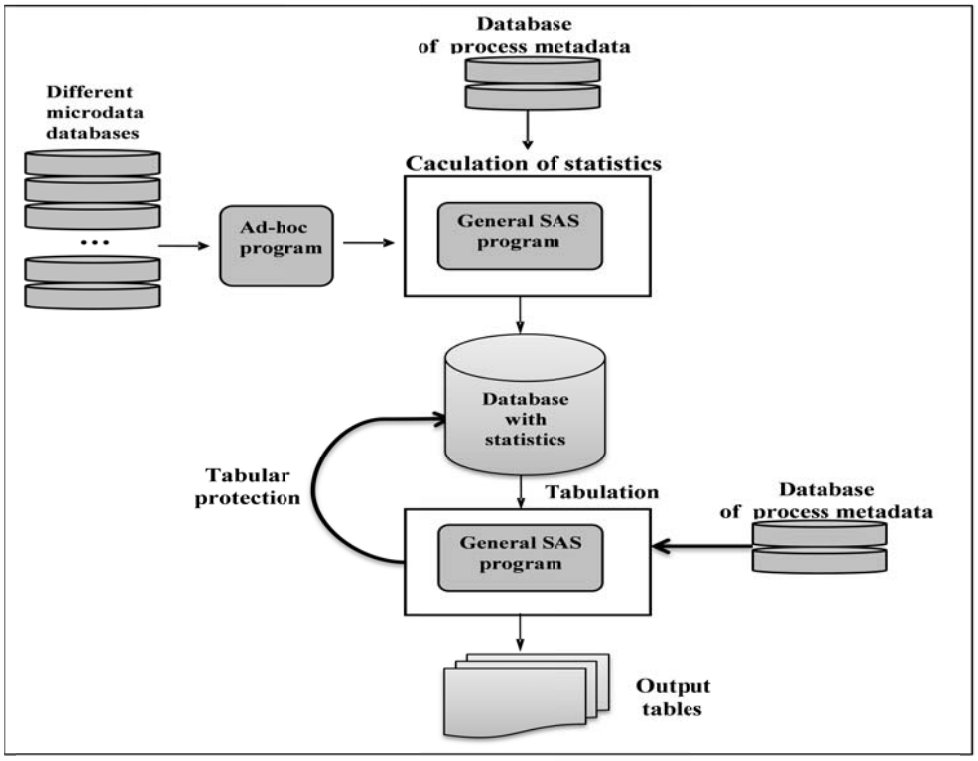


Figure 1: Schematic presentation of the functioning of the building block "data editing"

9. The system presented above is slightly different in the case of aggregation, application of SDC methods and tabulation. The input table for the calculation of statistics is also a SAS (work) table, which consists of all edited microdata; the output is written to one specific database, where all statistics with respect to the domains' values are listed. In case of applying the cell suppression SDC method, special formats of tables are created, which can be used with Tau Argus, software program designed to protect statistical tables. Tau Argus is one of the results of one of the EU projects, mainly developed at Statistics Netherlands. The confidential statuses are added to the database with statistics. After that output tables with

confidential pattern and denotations for lower degree of precision are created. One of the possible formats of output tables is also suitable as direct input in PX-Edit, which is our tool for publication on our data portal. A simplified schematic presentation of the functioning of the building block “tabulation” is presented in the following figure:



10. The whole metadata is linked to the input tables (there can be several input tables for one survey), which consists all the microdata needed for the individual part of the statistical process. The only exceptions are the data needed for the sampling error estimation (sampling rate for each of the stratum cells), which are linked to survey instance, and the code lists for the tabulation, which are linked to the survey.

III. Renewed system

11. The first version of the generic system uses general SAS programs, which are able to read the general rules for data processing (process metadata) and on the basis of these metadata create the desired output. It is a very open system, which allows that input microdata as well as process metadata can be stored in different environments. Such an open system is definitely highly flexible and provides a suitable tool for building up a statistical process.

12. However, there are also some obvious shortcomings of such an open system. These shortcomings are mostly related to the process metadata management procedures. As indicated, the database of process metadata has a strictly determined structure, but it can for each particular survey be placed in different databases and even in different environments (e.g. ORACLE, MS Access, SAS). In fact, for most of the so far included surveys the process metadata were stored inside the MS Access databases. The reason for this was mainly that subject-matter specialists, who are predominantly in charge of managing these metadata,

prefer this environment due to its simplicity and user friendliness. The consequence of such a practice is that the process metadata are at the moment scattered all over the different network directories in different Access databases.

13. The problem with the scattered system of process metadata is that it is impossible to create an effective general application for managing and controlling the inserted metadata. As it was pointed out in the analyses after the first period of the usage of the disintegrated system, the most problematic part was the significant number of errors in the process metadata. Since the fields for inserting rules were entirely open fields, most of these errors were errors in the syntax of the rules (e.g. bracket errors) or errors in consistency between rules and variables. All the building blocks in fact incorporate a certain number of checks which control the consistency of the provided metadata (e.g. check if the variable to be imputed is in the input data table), but all these checks can only be performed subsequently, during the execution of the process.

14. To enable the creation of a better system for process metadata management and navigation, we decided to perform a certain degree of the re-integration of the whole system. The aim of this re-integration is certainly not to build again the fully integrated system as initially designed, but to re-integrate only to such a level that would on one hand enable creation of the general management tool but would on the other hand keep the high flexibility of the system. The following re-integration actions were decided to be carried out:

a) To build one single, unique database of process metadata. This database would be created in ORACLE and managed by the .NET application, which would enable user friendly management of the process metadata.
b) To connect the system with the metadata repository, where the data on surveys and survey instances are stored.
c) To enable the application management to the access of the process metadata for each survey.
d) Some ad-hoc SAS programs that are “survey-dependent” (preparation of the input tables and writing back to microdata database) are still needed.

15. The architecture of the renewed system is presented in the following figure:

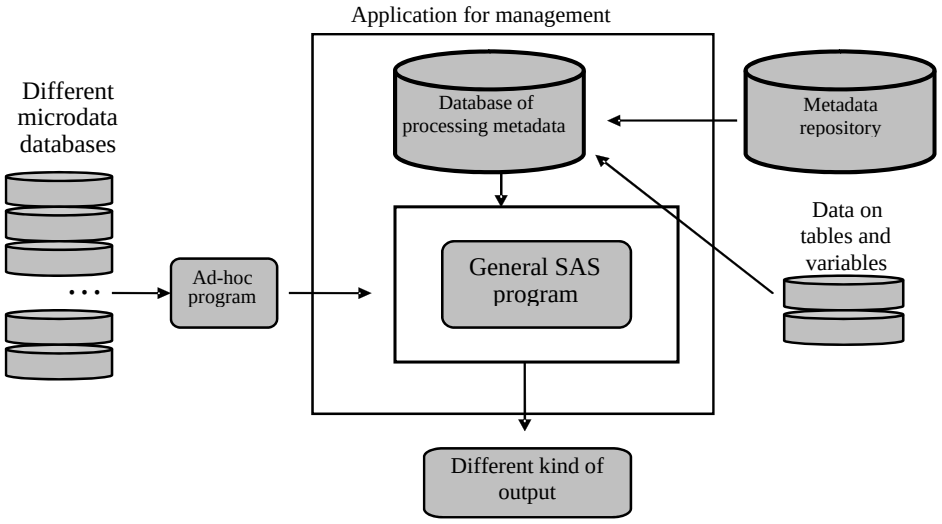


Figure 3: Schematic presentation of the renewed system

16. Development of the renewed system is carried through the project that was launched in 2012. SURS's employees from three different areas are included in the project team: experts from general methodology, IT experts and subject-matter specialists. Development work is focused on three main goals: to renew and upgrade the general SAS programs, to create a unique database of process metadata and to develop graphical interfaces for management of process metadata.

17. From the user's point of view the .NET graphical interfaces are the central point of the whole system. Through these interfaces the user selects the survey, selects the survey instance, inserts or edits the process metadata and also runs the particular part of the statistical process. This application, called the MetaSOP, is hence the key output of the project.

18. There are quite a large number of the graphical interfaces; here we present just one, which provides the user possibility of calculation of one type of the statistics.

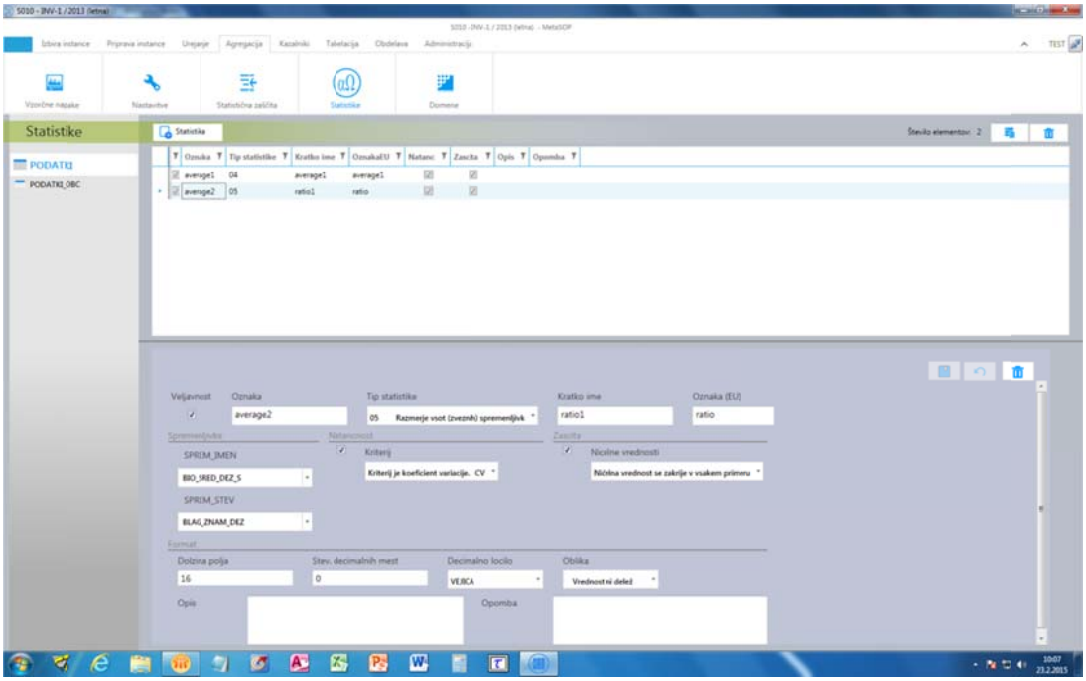


Figure 4: Graphical interface of the application MetaSOP

19. We have been developing our MetaSOP application very intensively over the last two years. For the building block "editing and imputation", the application has already been finished and for several surveys it has already (from beginning of 2015) been used in regular production. For the second part of the statistical process (SDC, sampling error estimation and tabulation) some finalisation is still needed to be done; therefore, the introduction into the production is planned for the second half of 2015.

IV. Challenges when a totally new system is introduced

20. When we started to introduce the new application (with the very new approach) to the regular production of the statistical surveys, we regularly collected the feedbacks from the survey methodologists that were faced with the new way of data processing. The main

advantages and main drawbacks of the new approach, as perceived by them, can be summarised as follows.

21. Main advantages:

a) The subject-matter personnel are much more independent of the IT department, which was previously in charge of the technical execution of the processes.
b) The rules for the data processing can very quickly be changed through the centralised system of process metadata. This makes the whole data processing cycle much more flexible.
c) Since the user can run the procedures several times in short time, it is now easier to check the feasibility of different methods for data processing.

22. Main drawbacks:

a) In the process of the insertion of the metadata expressions there is a high risk of syntax errors. As the consequence, the application cannot be executed or is executed with the wrong parameterization.
b) The subject-matter specialists need to learn some new skills, which is sometimes a problem in the reality of the very burdensome statistical production.
c) If an error occurs during the execution of the procedure, the technical staff must be contacted and if they are not available, the process execution can stop for some time.

23. Development of a totally new system for statistical production is certainly a big step forward for SURS. We firmly believe that the project outcomes will help us to build a new, modernised system of management of statistical data processing in different statistical systems. Movement from stove-pipe to centralised methodological and IT solution is the final goal of these developments. The central point of the renovated system is the metadata driven application, which is on one hand flexible in the sense that it can be plugged to different microdata environments, while on the other hand introduces very centralised management of the process metadata.

24. Introduction of such generic, MDD application for data processing unavoidably introduces certain changes also at the general, institutional level when design and implementation of the statistical surveys are concerned. Based on the experiences gained so far, the main changes can be summarised as follows:

a) There is essentially different distribution of work between subject-matter specialists, general methodologists and IT experts. With the old system, each subject-matter specialist had his or her “own programmer” and his or her “own general methodologists”, who used the specific instructions of the subject-matter specialist to design and implement ad-hoc processes for a certain survey. Now the general methodologists and IT experts act only as the “support team” in the case when certain error in the application occurs or the process doesn’t provide the expected results. This means that subject-matter specialists are now much more independent of the IT Department and the General Methodology Department.
b) Change in the role of subject-matter specialists in the statistical process also changed expectations of their skills and capabilities. It used to be expected that they have a very deep knowledge of the subject-matter and that they are capable

of providing the written instructions (in open form) for implementation of certain parts of the process (e.g. imputation, aggregation). Now it is expected that they are trained and educated to be able to write these rules themselves already in the form of mathematical-computer language.

c) The whole organization of work of the IT Department and the General Methodology Department will have to be changed from domain oriented to process oriented. This re-organisation means a significantly different general view to the institution's organisation and distribution of work and is therefore quite a challenge for the statistical organisation. SURS is at the moment facing the first stages of dealing with this challenge.

d) The above described re-direction from (specific) domain oriented to (general) process oriented production will have to be realized also at the level of the functioning of our IT and methodology experts. Developing and supporting of such generic applications require experts capable of thinking and operating at a much more general level, considering the execution of a certain survey just as one of the realisations of the general statistical process.

IV. Conclusion

25. The paper presents activities that have recently been carried out at SURS aiming to develop and implement generic IT solutions for data processing. With this development work we are in fact trying to solve a 'classical' problem of transition from the stove-pipe oriented production to the more integrated processing systems. The development activities are oriented toward more standardised methodological as well as more standardised and particularly more generic IT solutions.

26. The development was carried out in two phases. During the first phase the first prototype solutions, which already encompassed all the key features, were developed. The key elements of the prototype solutions were small generic, metadata driven SAS programs, developed to cover the particular part of the statistical process for a wide range of different surveys. These solutions have been successfully used in several large and demanding surveys (e.g. the 2010 Agriculture Census, the 2011 Population Census) and have already significantly contributed to the improvements of our process in the sense of efficiency, harmonisation and standardisation.

27. Although these applications used general SAS programs based on general metadata tables, there were still a lot of features which characterize these solutions as local solutions. Therefore, in 2012 SURS launched a large project with the goal to upgrade the existing solutions and to build one global solution for the data processing. In the paper we presented the main features of this new solution and also described which changes would be caused by the introduction of such a generic tool in the overall organisation of the statistical process at the institutional level. Successful introduction of these institutional changes is certainly one of the main challenges of SURS for the following years.

References

Dolenc, D., Krek, M., Seljak, R. (2011), "Editing Process in the Case of Slovenian Register-based Census", paper presented at the UNECE Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9-11 May 2011)

Seljak, R. (2009), “Integrated statistical systems and their flexibility – How to find the balance?”, Presented at the NTTS conference, Brussels, Belgium, 5-7 March, 2013

Seljak, R., Blazic, P. (2011), “Sampling error estimation – SORS practice”, Presented at the 2nd European Establishment Statistics Workshop, Neuchatel, Switzerland, 12-14 September, 2011

Seljak, R. (2014), “Metadata driven application for data processing – from local toward global solution”, paper presented at the UNECE Work Session on Statistical Data Editing, Paris, France, 28-30 April 2014

Van der Veen, G. (2007). Driving forces for changing Dutch statistics. Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, Volume 24, Number 1-2 / 2007.

Zeila, K. (2004). Metadata Driven Integrated Statistical Data Management System. Presented at Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS): Geneva, 17-19 May 2004.

Smukavec, A. (2013), “Metadata driven application for aggregation and tabular protection”, paper presented at the UNECE Work session on statistical data confidentiality, Ottawa, Canada, 28 - 30 October 2013.