Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (v): International collaboration and processing tools

Metadata driven application for data processing – from local toward global solution

Prepared by Rudi Seljak, Statistical Office of the Republic of Slovenia, Slovenia

I. Introduction

1. Statistical data processing has always been a demanding, time consuming and consequently quite 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 and the accuracy and reliability of the results is a key issue. The consequence of this fact is that a lot of resources have to be spent for the data processing, especially for the data validation, cleaning and statistical editing. 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 the statistics of high (or at least sufficient) quality with the significantly reduced resources.

2. To overcome or at least reduce the gap between the above mentioned demands, in the 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 six years 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). From then on these applications were gradually developed and supplemented and have so far been successfully used in several large and demanding surveys, such as the 2010 Agriculture Census and the 2011 Population Census. Although these applications 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, in 2011 SURS launched a project aiming at upgrading the existing solutions and building one global solution which would cover all before mentioned parts of the data processing and which could easily be used for most of the statistical surveys.

4. The paper presents the main characteristics of the developed tools with the especial emphasis on the development of the generalised tool and how the introduction of such a tool can change the design of the whole statistical process. The technical details of the new tool were already described in previous papers (see Seljak 2009; Seljak 2011 and Seljak and Blazic 2011), so here the focus is on a more general point of view of the introduction of such a tool.

II. Generalised solutions – main characteristics

A. Building blocks

5. Modernization and standardization of the SORS production system is based on the development of small, generic solutions, which are 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, which we also call the building blocks, provide the generic software solution for the certain part of the statistical chain and are designed in a way that they can act rather independently of each other. The main features of these building blocks could be summarised as follows:

(a) They are designed on the basis of harmonized, transparent and widely accepted methodological principles, which had been determined before the actual creation of the particular building block.

(b) They are opened to such extent that it is not required that all the data inputs come from one unique, comprehensive database. In other words: 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.

(c) They are designed as fully metadata driven (MDD) systems, meaning that information which determines the parameters for the execution of the processing for the concrete survey and concrete reference period are provided outside the core computer code. No information referring to concrete survey execution is incorporated into the general program code but is provided by the subject-matter personnel through the special metadata tables.

(d) The process metadata can also be provided in different databases in different environments, but each of these (metadata) databases must follow the strict rules of its structure (tables and variables).

6. The main object that presents the input in the building block is a table, where all the microdata that are to be processed are captured. In the current execution of the system, this table has to be a SAS (work) table. Also the output of the building block is a SAS table. Therefore, we always need a small ad-hoc program which prepares such a table and an ad-hoc program which transfers the output table back to the database. A simplified schematic presentation of functioning of such a building block is presented in the following figure:

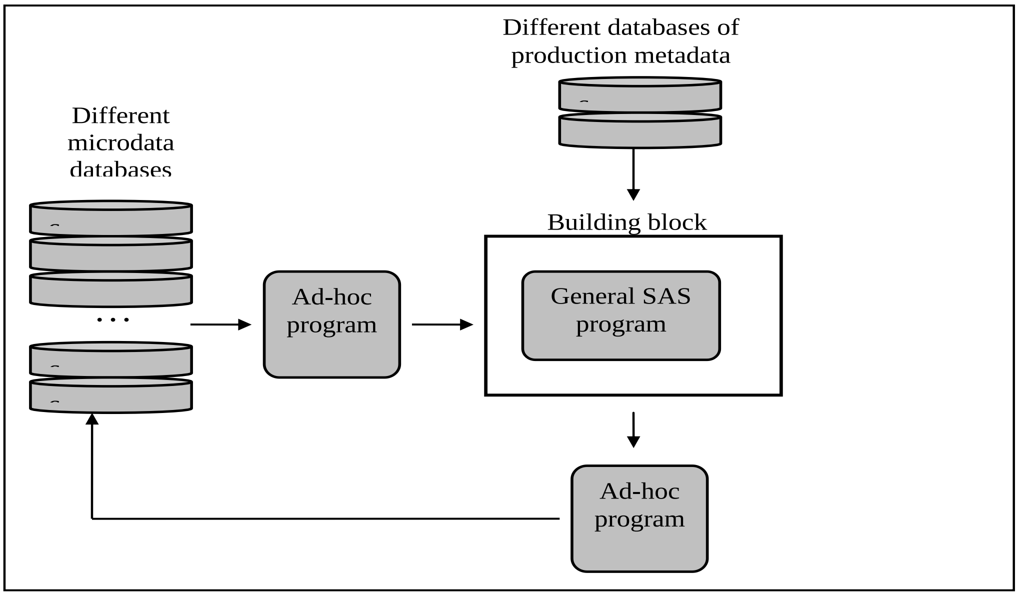

Different databases of production metadata
Different microdata databases
Building block
Ad-hoc program
General SAS program
Ad-hoc program

Figure 1: Schematic presentation of functioning of the building block

7. One building block usually covers only a small part of the statistical production chain. Therefore, the inputs and outputs of these components must in the final stage be linked together. A simplified schematic presentation of the whole process is presented in the following figure:

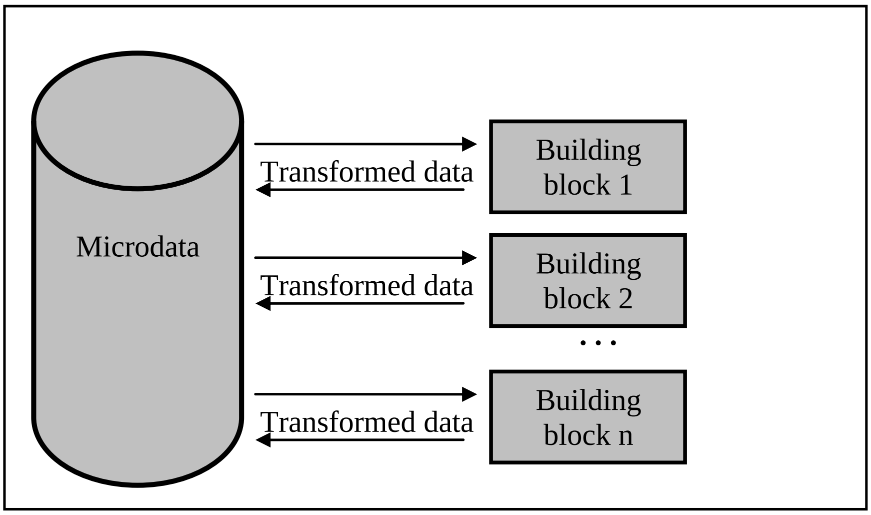


Figure 2: Linking inputs and outputs of the building blocks

B. Process metadata - example

8. As it was already explained, the technical details of the system were already described in the previous paper. Here we only provide a short illustrative example of the usage of the process metadata for one of the imputation rules. Suppose that we want to impute the missing values for three variables Occupation (OCC), Education (EDUC), and Activity status (ACT) in the micro-data table called TABLE1. We decide that we will use the hot-deck method in three steps. In the first step we will impute the missing values for the OCC variable by searching for the donor inside the same municipality (MUN). Since we will require a minimum of 10 donors in the imputation cell, some missing values will probably not be imputed in the first step. Therefore, we will use another step in the process, where we will search for the donor inside the region (REG) and decrease the minimum number of donors to 5. The matching variable for the hot-deck method will in both cases be age of the person (AGE). When the missing values for the OCC variable are imputed, we will then use this variable for the purposes of the imputation of EDUC and ACT variables. Both variables will be imputed in the third step with the same (version of the) method where donors will be searched inside the same occupancy class and age will again be used as the matching variable. The basic metadata for this procedure is given in the following metadata table:

Table 1: Metadata for imputation procedures

Table	Variable	Cond_imput	Cond_donor	Step	Method
TABLE1	OCC	OCC Is Null	Not (OCC Is Null)	1	HD1
TABLE1	OCC	OCC Is Null	Not (OCC Is Null)	2	HD2
TABLE1	EDUC	OCC Is Null	Not (OCC Is Null)	3	HD3
TABLE1	ACT	OCC Is Null	Not (OCC Is Null)	3	HD3

Short description of the fields:

Table:	Name of the input SAS table which is to be processed
Variable:	Name of the variable which is to be imputed
Cond_imput:	Condition which determines for which units the imputation procedure will be performed
Cond_donor:	Condition which determines which units can serve as the suitable donors
Step:	Step of the imputation process
Method:	Version of the hot-deck method

9. The method field needs some further explanations. In fact, users can create an arbitrary number of their own versions of the hot-deck method. Since each newly created version of the method can be used to impute several variables, the rules for these versions are given in a separate table:

Table 2: Metadata for imputation methods

Method	Strata	Match_Var	Min_donor
HD1	MUN	AGE	10
HD2	REG	AGE	5
HD3	OCC	AGE	5

Short description of the fields:

Method: Version of the hot-deck method
Strata: Stratification variables that determine the definition of imputation cells. If several variables are included, they should be separated by ","
Match_var: Matching variable
Min_donor: Minimum number of donors inside the imputation cell

III. Building a global solution

A. Why the new application is needed

10. The described system uses three basic objects as inputs: tables, variables and rules. As it was explained earlier, the concretisation of these abstract objects can be quite arbitrary and is a subject of particular parameterisation. Such an open system is definitely highly flexible and provides a suitable tool for building up a statistical process. However, there are also some obvious shortcomings of such an open system. These shortcomings are mostly related to the process metadata management procedures. As we already indicated in the previous sections, 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.

11. The problem with such a 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 'building-block system', the most problematic part was the significant number of errors in the process metadata. Since the fields for inserting rules are at the moment 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 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.

12. In order to enable the creation of a better system for process metadata management and navigation, the new project was launched, aiming at creating the new platform which would integrate the scattered parts to such a level which 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 actions were decided to be carried out during the project:

(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 connect the system with the metadata repository, where the data on surveys and survey instances are stored.
(d) To enable creation of the process metadata and launching of the process from a single, central point.

A. The new SOP[1] application

13. The new application is at the moment still in the development phase. It is planned that the part of the application which aims at covering the data editing part of the process would be introduced in regular production in the second half of 2014. The second part of the application, which would cover the part of aggregation, tabulation, standard error calculation and data disclosure control, is planned to be finished in 2015.

14. The application is developed by combining three IT environments:

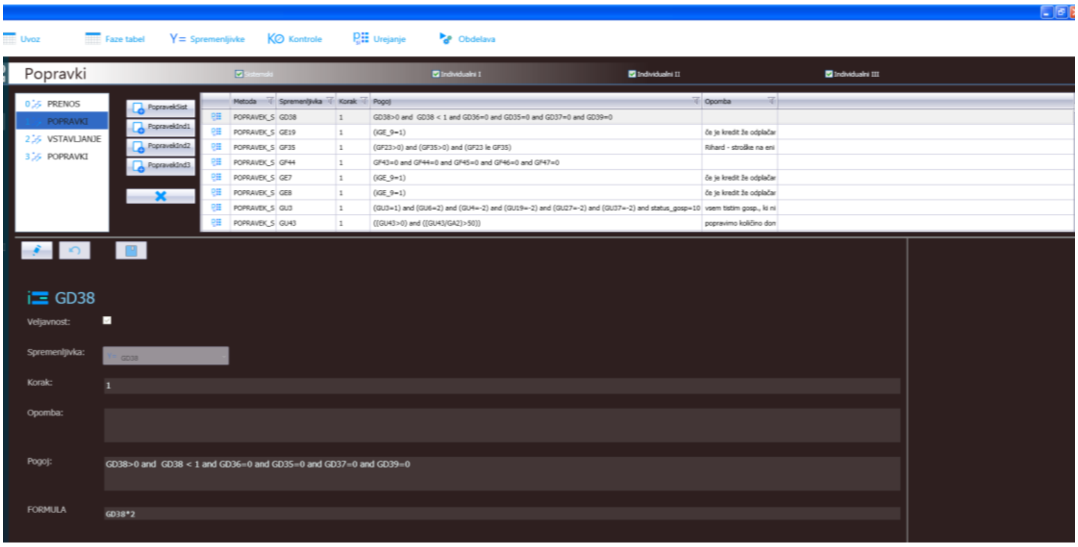
(a) ORACLE database for unique, central database of process metadata
(b) SAS macros as general programs for data processing
(c) .Net environment to build the graphical interfaces for management of the whole system

15. As far as the users (domain statisticians) of the application are concerned, they will only face the .Net application. The remaining two systems will be “hidden” in the background and will only be accessed by developers and administrators.

16. There are quite a large number of the graphical interfaces, but roughly we can divide them into the following three sets:

(a) Interfaces for selection of the survey and survey instance
(b) Interfaces for management of the process metadata
(c) Interfaces for running the processing (running the SAS macros)

17. We demonstrate the set of graphical interfaces by presenting two of them. The first one is the graphical interface aimed at facilitating creation and editing of the rules for deterministic data corrections. Figure 2 presents the interface for running the processes that were for a particular table determined previously.



Metoda	Spremenljivka	Korak	Pogoj	Opomba
POPRAVEK_S	GD38	1	GD38>0 and GD38 < 1 and GD36=0 and GD35=0 and GD37=0 and GD39=0	
POPRAVEK_S	GE19	1	(GE_9=1)	če je kredit že odplačar
POPRAVEK_S	GF35	1	(GF23>0) and (GF35>0) and (GF23 le GF35)	Rihard - stroške na eni
POPRAVEK_S	GF44	1	GF43=0 and GF44=0 and GF45=0 and GF46=0 and GF47=0	
POPRAVEK_S	GE7	1	(GE_9=1)	če je kredit že odplačar
POPRAVEK_S	GE8	1	(GE_9=1)	če je kredit že odplačar
POPRAVEK_S	GU3	1	(GU3=1) and (GU6=2) and (GU4=-2) and (GU19=-2) and (GU27=-2) and (GU37=-2) and status_gosp=10	vsem tistim gosp., ki ni
POPRAVEK_S	GU43	1	((GU43>0) and ((GU43/GA2)>50))	popravimo količino don

GD38

Veljavnost: ☑

Spremenljivka: GD38

Korak: 1

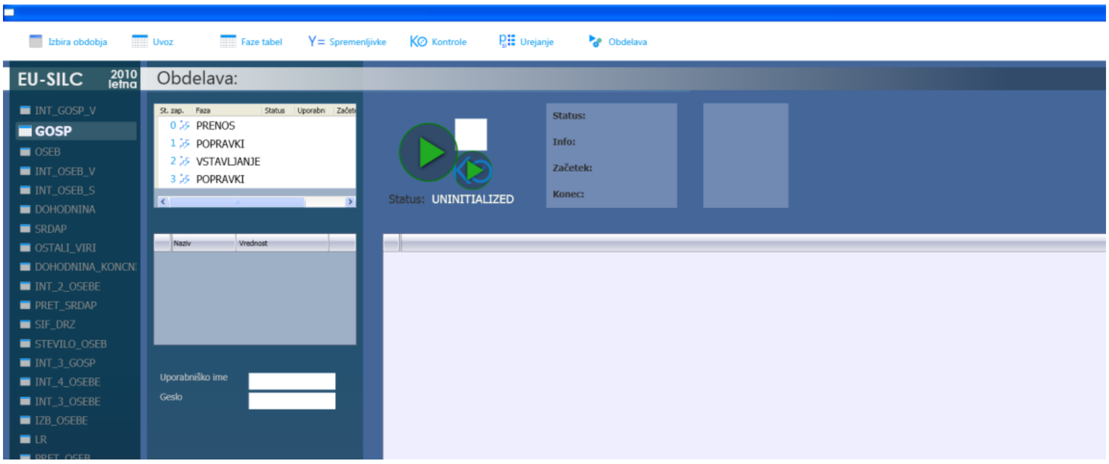
Opomba:

Pogoj: GD38>0 and GD38 < 1 and GD36=0 and GD35=0 and GD37=0 and GD39=0

FORMULA GD38*2

Figure 3: Interface for deterministic corrections rules

[1] SOP is Slovenian acronym for Statistical Data Processing



St. zap.	Faza	Status	Uporabn	Začet
0	PRENOS			
1	POPRAVKI			
2	VSTAVLJANJE			
3	POPRAVKI			

Status: UNINITIALIZED

Status:
Info:
Začetek:
Konec:

Uporabniško ime
Geslo

Figure 4: Interface for execution of the processes

IV. New application and statistical process

18. Introduction of the before presented generic, MDD application for data editing into the statistical process unavoidably introduces certain changes also on 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 IT specialists, general methodologists and IT experts. With the old system, each subject-matter statistician had his or her “own programmer” and his or her “own general methodologists” (specialist for data editing methods), which 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 from the IT Department and the General Methodology Department.

(b) Change in the role of subject-matter statisticians in the statistical process also changed expectations of their skills and capabilities. Before it was 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 creation of certain parts of the data editing process (e.g. edit checks). Now it is expected that they would be trained and educated to be able to write these edit checks themselves already in the form of mathematical-computer language[2].

(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 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. Conclusions

19. The paper presents activities carried out by SURS in the recent years in order to develop and implement generic, MDD IT solutions for data processing. It is in fact a classical problem of transition from the stove-pipe oriented production to the more integrated processing systems. SURS developed the

[2] This language is in our case the SAS program code syntax

generalized MDD applications which support certain parts of the statistical process (e.g. deterministic corrections, imputations, standard error calculation) several years ago. These applications have already been successfully used in several large and demanding surveys (e.g. the 2011 Population Census) and have already significantly contributed to the improvements of our process in the sense of efficiency, harmonisation and standardisation. Although these applications use general SAS programs, based on general metadata tables, there are still a lot of features which characterize these solutions as local solutions. Therefore, SURS is at the moment implementing a project which aims at upgrading the existing solutions and building one global solution for the data processing part of the process. The paper 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.

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, Slovenia (Ljubljana).

Seljak, R. (2009), “New Application for the Slovenian EU-SILC Data Editing”, Presented at the UNECE Work Session on Statistical Data Editing, Neuchatel, Switzerland, 5-7 October, 2009

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