REPUBLIKA SLOVENIJA
STATISTIČNI URAD RS

Statistical data processing with R - metadata driven approach

Rudi Seljak
Jerneja Pikelj

Statistical Office of the Republic of Slovenia

7. April 2016

Contents

- Statistical data procesing at SURS
- General solutions - main characteristics
- Metadata driven principle for data validation with R
- Conclusions

Statistical data procesing at SURS

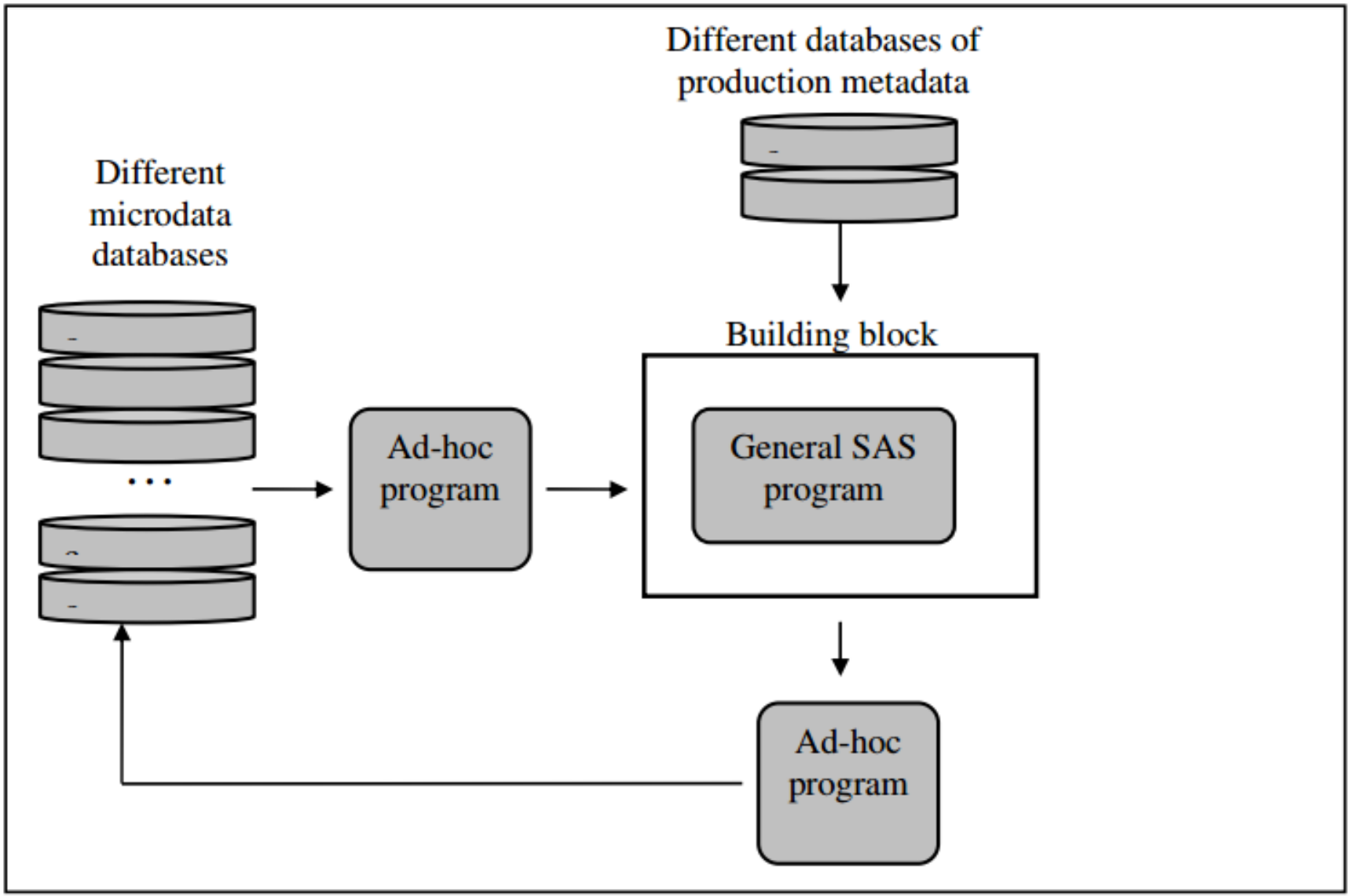
- Rationalization of statistical processes
- The need for transition:
  - from custom made solutions for surveys (stove-pipe approach) to generalized process solutions
  - from domain oriented to process oriented solutions
- Main goals of the project:
  - repetability of the data production process
  - transparency of the data production process

General solutions - main characteristics

Modernization and standardization of the system of statistical production at SURS is based on the development of small generic solutions - building blocks.

All the parameters for the particular survey and the particular reference period is located in special database, called metadata database.

Schematic presentation of the building block



Metadata driven principle for data validation with R

Short ilustrative example on the data from Monthly Statistical Survey on Earnings Paid by Legal Persons.

- Ad-hoc program
- Metadata
- General program
- Results of data validation

Ad-hoc program

- Input work table with microdata about earnings paid out by around 50.000 legal persons of the public and private sector
- Key variables:
  - BRUTO_PLACA - gross earnings paid out for the reference month
  - NETO_PLACA - net earnings paid out for the reference month
- We save our microdata table into temporary table data.

Metadata - 1

- In our case the metadata table is constructed in Excel, but in general it can be located in any other database, which can be connected to R, for example MS Access or ORACLE.
- Structure of metadata table for data validation:
  - LC_LABEL - Label of the logical check. It has to begin with LK.
  - ERROR_DESCRIPTION - Description of the error.
  - CONDITION - Condition which determines our check.
  - ERROR_TYPE - Type of an error e.g. error, warning or other options.
  - VALIDITY - Validity for the specific check. If the value is zero, the check will not be executed..

Metadata - 2

TABLE	LC_LABEL	ERROR_DESCRIPTION	CONDITION	ERROR_TYPE	VALIDITY
DATA	LK002	Error, if the net wage is greater than gorss wage.	BRUTO_PLACA<NETO_PLACA	ERROR	1

```
checks<-read.xlsx("Metadata.xlsx", sheet = 1, startRow = 1, colNames =
TRUE,skipEmptyRows = TRUE, rowNames = FALSE)
  > checks

     TABLE  LC_LABEL  ERROR_DESCRIPTION
[2]  data   LK002     Error, if the net wage is greater than gorss wage.

     CONDITION                    ERROR_TYPE   VALIDITY
     BRUTO_PLACA<NETO_PLACA       ERROR            1
```

General program - 1

- First, we create new variable R_st, where the programing code for logical checks is defined.

```
 R_st <- paste(checks$TABLE,"$",checks$LC_LABEL,"<- ifelse(",
checks$CONDITION,",1,0)",sep= "")

 > R_st
[2] "data$LK002<- ifelse(BRUTO_PLACA<NETO_PLACA,1,0)"
```

General program - 2

- Attach the variables in the workspace

```
attach(data)
```

- Execute the code, defined in R_st

```
for (i in 1:length(R_st)){
  eval(parse(text=R_st[i]))
}
```

Results

- Output by ID

	MAT_ST	LK002
1	xxxxxxxxx1	0
2	xxxxxxxxx2	0
3	xxxxxxxxx3	0
4	xxxxxxxxx4	0
5	xxxxxxxxx5	0
6	xxxxxxxxx6	0

- Summarised table

	nr_down	ERROR_DESCRIPTION
LK002	**0**	Error, if the net wage is greater than gorss wage.

Conclusions

- Advantanges and disadvantages of using R
- The importance of the data validation
- Opportunities of using R in the future

Thank you for your attention!