Rudi Seljak
Statistical Office of the Republic of Slovenia, e-mail: rudi.seljak@gov.si

Abstract
In the paper we use experiences from Statistical Office of the Republic of Slovenia to discuss some questions related to the problem of transition from the classical stove-pipe oriented production to the integrated processing systems. Although such orientation certainly presents the right direction for the future development, there are many pitfalls related to such transition. We describe our redirection from the plans for comprehensive, fully integrated system to a revised system with the certain degree of disintegration, which turned out to be more flexible and effective. The theoretical discussion is supplemented with the example from the area of data editing and imputation.

Keywords: metadata driven systems; process metadata; generic applications

1. Introduction
Production of accurate and reliable statistical results 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. On the other hand, the official statistics usually targets the large populations and complex phenomena inside these populations, which makes the whole task even more demanding. With the constant pressure for budget cuts, the official statisticians are hence more and more facing the challenge of producing the statistics of high (or at least sufficient) quality with the significantly reduced resources.

One fact that certainly acts in favor 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.

It was somewhere at the beginning of the new millennium when the ideas of fully integrated systems for the statistical production became very popular (see Van der Veen 2007 and Zeila 2004). The general idea of these systems was to build the comprehensive
IT environment which would support the whole statistical production cycle and could, with a minimum number of adjustments, serve for the purposes of all (or at least of most) of the different statistical surveys. These systems would with one “big leap” enable transition from the classical stove-pipe oriented production to the modern generalized and standardized statistical production. Now after more than a decade of research and development in this direction it is perhaps time for critical reflection of these ideas and critical analyses of the concrete results that were achieved by following these ideas.

This paper aims at giving a modest contribution on this topic by describing the concrete experiences from the Statistical Office of the Republic of Slovenia (hereinafter SURS). In the first part of the paper we briefly describe our attempts towards the fully integrated system and explain the main obstacles that we faced during these attempts. The central part of the paper is then dedicated to the description of the main features of the adjusted strategic approach which was defined by using the previous experiences and is based on a certain degree of disintegration of the whole system. We give some conclusions and directions for future developments in the last part of the paper.

2. Integrated system at SURS

In 2007 SURS started a large project which aimed at creating a complex, fully integrated system which would be able to cover the whole production cycle for a vast majority of statistical surveys. This integrated system was planned to develop and integrate the following main elements:

- Renewed and enhanced Statistical Business Register, which should become a central point for storing and maintaining the exhaustive set of information for all the statistical units targeted by our statistical surveys.
- A single and commonly used database of questionnaires, questions, concepts and variables.
- New infrastructure for electronic data collection and enhanced system for respondent’s management.
- Generic, metadata driven system for the whole cycle of the statistical data processing.

The project lasted two years and required a lot of internal as well as some external (outsourced) resources. At the end of the project we had to admit that the results of the project only partially met the high expectations. There were some useful tools created during the projects, which could be (with some additional input) well used in the statistical production. However, the project certainly failed in the aim of creating a fully integrated system which would link together all the particular tools that were provided through the project. There were several reasons for only partial success of project implementation. The most obvious, as pointed out by the posterior analyses, were:

- It was too complex a project with regards to the financial and human resources available for its implementation.
- Two-year duration of the project was certainly too short a period to create such a complex system.
Some elements of our statistical infrastructure were not really in the state which would enable transition to such a complex system.

Besides above stated reasons, in fact “the usual suspects” for the insufficient project success, there was also one consideration which was more and more frequently raised towards the end of the project: “Are such fully integrated systems really the best development orientation?” In other words: “Do such integrated systems really assure the modernization towards the flexible and (cost) effective statistical production system?” In the following sections of the paper we will try to discuss this issue in greater detail. On the basis of our experiences, we will show what were our main reasons to give up the idea of a fully integrated system and turn to some slightly different solutions, which still keep some important features of the “old system” (e.g. metadata driven approach), but certainly use a certain degree of disintegration which will in our opinion certainly contribute to larger flexibility and generality of the system.

To make the discussion “reasonably complex”, we mostly focus on that part of statistical production in which the data are transformed at the micro level. Usually this part of the process is known as “editing & imputation” procedures (hereinafter E&I procedures). We will use this part of the statistical process to show the main differences in both approaches.

3. Towards a disintegrated system

3.1 The case of E&I procedures

The E&I procedures were in the integrated system planned to be based on the following assumptions:

- There is a single, unique database of microdata for all the statistical surveys which would be included in the system. The only way to make such a database feasible is to organize the data according to the “normalized data model”, meaning that each value of a particular variable for a particular unit is written in a separate record and the name of the variable is just one of the attributes of the record.
- Each variable (basic or derived) which is supposed to be used in the E&I procedures should be previously registered in the register of questionnaires, questions, concepts and variables, which is again a single and unique database for all the statistical surveys in the system.
- The reference metadata for the survey should be available and accessible through the usage of the single, unique metadata database (metadata repository). Such a repository in fact operationally already existed before the beginning of the project.
- Processing metadata (rules for execution of the E&I procedures) are recorded in a single, unique database of the processing metadata.
- There is a generic (SAS) program, which is able to carry out the actual E&I procedures on the basis of the information available in the above listed databases.

The architecture of the planned system is presented in the following figure:
As can be seen from the description and enclosed figure, the main problem of the system was that it required several large and rather complex systems to be successfully integrated, to then communicate smoothly and to provide the user friendly way to manage and to navigate it. In fact the main conclusion that was drawn out of the analyses of the project results was that the operability of the whole system depends on too many sub-systems and that for the more effective further development we certainly need to disintegrate it into smaller components.

Therefore, at that point we decided to build up a new strategy for modernization and standardization of our production system. We decided that we will break the integration system 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, which we also call the building blocks, should provide the generic software solution for the certain part of the statistical chain and should be designed in a way that they can act much more independently than the procedures in the (previously planned) integrated system. The main features of these building blocks could be summarized as follows:

- 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.
- They should be opened to such extent that it would no longer be required that all the 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 of the data.
- They are still designed as fully metadata driven systems, meaning that information which determines the parameters for the execution of the processing for the concrete survey and concrete reference period should be provided outside.
the core computer code. No information referring to concrete survey execution should be incorporated into the general program code but should be provided by the subject-matter personnel through the special metadata tables.

- 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).

The main object that presents the input in the building block is a table, where all the microdata, which aimed at being 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:

![Schematic presentation of functioning of the building block](image)

Figure 2: Schematic presentation of functioning of the building block

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:
To deal with the E&I procedures, three building blocks were developed, each of them responsible for the following procedure: Logical controls; Deterministic corrections of erroneous data\(^1\); Data imputation for missing and inconsistent data. These building blocks can then be linked together in the E&I process by using arbitrary sequence of the basic sub-processes. The typical E&I process can hence begin with the deterministic corrections, followed by imputations and then by one more (final) correction. Each version of the data is controlled by using the set of pre-defined logical controls. The process for this particular case is presented in the following figure:

\(^1\) With the deterministic corrections we here refer to the simple systematic corrections, expressible in "IF-THEN" form (e.g. IF V1<0 then V1=0), as well as individual corrections which are performed at the level of a particular variable for a particular unit.
input in the subsequent component. Namely, most of the process metadata for corrections and imputations should be created on the basis of the results of the logical controls.

3.2 Process metadata

The whole metadata driven system for the building blocks was already thoroughly described in some previous papers (see Seljak 2009; Seljak and Blazic 2011), so we here provide only a short illustrative example. As mentioned, the whole system is to a very large extent based on the process metadata. To illustrate the structure and syntax of these metadata, we present a simple example of process metadata for one of the imputation rules. Suppose that we want to impute the missing values for the variable Occupation (OCC) in the table called TABLE1. We decide that we will use the hot-deck method in two steps. In the first step we will search 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). The basic metadata for this procedure are given in the following metadata table:

<table>
<thead>
<tr>
<th>Table</th>
<th>Variable</th>
<th>Cond_imput</th>
<th>Cond_donor</th>
<th>Step</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE1</td>
<td>OCC</td>
<td>OCC is Null</td>
<td>Not (OCC is Null)</td>
<td>1</td>
<td>HD1</td>
</tr>
<tr>
<td>TABLE1</td>
<td>OCC</td>
<td>OCC is Null</td>
<td>Not (OCC is Null)</td>
<td>2</td>
<td>HD2</td>
</tr>
</tbody>
</table>

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

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>
<thead>
<tr>
<th>Method</th>
<th>Strata</th>
<th>Match_Var</th>
<th>Min_donor</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD1</td>
<td>MUN</td>
<td>AGE</td>
<td>10</td>
</tr>
<tr>
<td>HD2</td>
<td>REG</td>
<td>AGE</td>
<td>5</td>
</tr>
</tbody>
</table>

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

4. Reintegration of the system

Using the highly abstractive description of the disintegrated system, we can claim that there are only three basic objects that this system is based on: tables, variables and rules. Such a highly generalized and open system is surely highly flexible and provides a suitable tool for building up a statistical process. However, there are also some quite obvious shortcomings of such an open system. These shortcomings are mostly connected 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 the fact that subject-matter specialists, who are predominantly in charge of managing these metadata, prefer this environment due to its simplicity and user friendliness.

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 disintegrated 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.

In order 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 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 re-integration actions were decided to be carried out:

- 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.
To connect the system with the metadata repository, where the data on surveys and survey instances are stored.

To enable the management application access to data on variables in incoming tables. These data can enter the application through two different channels: from the register of variables, if the survey has registered its variables there, or from some other location where the data tables structures are stored.

A simplified schematic presentation of the planned re-integrated system is presented in the following figure:

![Schematic presentation of the reintegrated system](image)

Figure 5: Schematic presentation of the reintegrated system

With this system we add two additional basic objects to the abstract description model: surveys and survey instances. Information on these objects is now exclusively available through the metadata repository, meaning that the basic condition to insert a certain survey instance into the system is that it is already registered in the metadata repository. On the other hand, we still kept the microdata in different databases. As we already described, the only condition for these databases is that they follow some basic general rules.

The partly re-integrated system is at the moment in the development phase and currently represents one of the most important development challenges for SURS. We plan to put the new system into production in 2014.

5. Conclusions

In the paper we gave an overview of the recent developments at SURS in the field of generalization and modernization of the statistical process. As in most of the NSIs, the main goal of these developments is to replace the classical stove-pipe oriented production
with some more effective production system which would be based on generic software solutions. We used our own experiences to critically discuss the idea of a fully integrated system and showed how in our case this idea was replaced with a slightly different approach, which still retains many important features of the initially planned system but is built on a certain degree of disintegration. The following main conclusions could be drawn from our experiences:

- The “big-bang approach” of moving directly from the classical stove-pipe system to the fully integrated system turned out not to be the best approach.
- The development of smaller generic solutions, called the building blocks, which could be easily plugged into the production system and therefore enable the successive transition to more generic solutions turned out to be a more effective approach.
- After creation of a certain number of building blocks and their successful testing in the regular production, the whole system should be re-integrated to a certain level again.
- The step-by-step approach is preferable to the “big-bang leap” plan. It is important that after each stage of the development the outcomes are critically and thoroughly analysed. The following action plan should be setup on the basis of the results of these analyses.

References

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

