The use of metadata to manage data processing processes and the definition of data validation rules

Keywords: validation, metadata, VTL, data processing, SQL, data processing system, software tools, validation language

1. INTRODUCTION

Conducting statistical surveys requires performing a number of specific steps beginning from collecting data, through processing them to preparing results. Each step is a process. The processes are dependent on each other and must be performed in a specific sequence.

The large number of such processes and their interconnectedness leads to difficulties in determining what processing step we are currently in and what the next step should be. The solution is to describe these dependencies in the form of metadata and to manage them appropriately.

One of such processes is the data sets validation, which consists data validation rules. Rules are created by experts in a given topic and then implemented in the data processing system by technical persons. Such implementation is not always trivial. The solution is to describe the rules in the form of notations that can be automatically translated into an executable form. Such records may be stored in the form of metadata and used by the data processing system.

2. METHODS

The following metadata structure was used to manage the processes:

- Dataset - represents a single data set (table in the MS SQL database or file (XML, DBF, XLS, XLSX, TXT, CSV)
- Period - the time interval assigned to process implementations and data sets.
- Data collection - represents a pair: a data set with a selected period. A single data set can contain several data collections.
- Process implementation - indicates an object that performs a certain process of processing statistical data.
- Process - is a representative of the process implementation for the selected data processing period. It is a single task of data flow / processing. Unlike its implementation, the process always applies to one period.

The state of the processed data is described by the statuses. Each data collection has a group of statuses.

The data collection and process are treated as individual objects, between which dependencies are determined. The data collection can be a source or target collection for a specific process. The process may be associated with several source collections and several target collections. These collections have certain statuses. This process can only be started if all data collections associated with it have statuses that meet the requirements defined in

the metadata. Each status of the source and target collections described in the requirements must assume one of the specified values. The process during execution can change statuses of the source and target collections associated with it directly.

On the basis of requirements and statuses, the order of processes and their dependencies are determined.

This mechanism allows to:

- manage the order of executing processes
- block the possibility of running multiple instances of the process at the same time
- manage the timeliness of the data stored in the collections
- track other important data properties (eg quality, availability, completeness)

The data validation processes are described in the metadata system. The individual error checks are described in the form of validation rules.

The rule consists of:

- Symbol – the unique identifier of the rule
- Data set - a set of data on which validation is performed
- Description of the incorrect situation - detailed explanation of the error in the data
- Error message - error message displayed in the validation report
- Technical notation - definition of validation in a form that can be translated it to the executable form for the data processing system (MS SQL). The following notations are allowed:
    - SQL - SQL query that returns a list of row identifiers
    - Where SQL - simplified SQL notation in the form of the where section of the SQL query
    - VTL - rule written in the the VTL language [1]

In addition to data validation rules, the system allows the definition of data editing and imputation rules.

3. RESULTS

Management of data processing processes is carried out in a special subsystem of the data processing system.

The data processing system provides users a tool called Processes manager. It shows data sets, data flows and processes in graphical way. If the process meet execution requirements user can be execute it. With this tool user see statuses of all processes and data sets. It’s also show properties of this objects. User can also look at content of the data sets.

Data validation is defined as processes operating on specific data sets. In addition, the process management system describes data sets containing validation rules. This allows to describe the rules synchronization processes and translate them to an executable form.

Translation of the validation notation to executable form is performed by a special service that uses partial VTL to SQL translator. The result of the translation is saved in the validation metadata.

4. CONCLUSIONS

The use of process management based on the metadata allows for easy manage data flows and document them. It also does not allow execute processes unnecessarily.

Defining validation rules in the form of special notations reduces the time between the rule definition and its implementation. It also reduces errors caused by communication between the technical and substantive persons.

REFERENCES

[1] VTR Reference Manual. available at: https://sdmx.org/?page_id=5096