A new international standard for data validation and processing

Marco Pellegrino (marco.pellegrino@ec.europa.eu)

**Keywords**: Data validation, transformation, open standards, SDMX, GSIM

1. **INTRODUCTION**

In the European Statistical System (ESS), Data validation is a critical issue which faces a number of problems due to the specificities of the production process. Data collection, processing and compilation are done by Member States (MS), while more processing steps and final dissemination at European level are performed by Eurostat. The role of Eurostat is critical for spotting any error in the figures transmitted, but in many cases Eurostat and MS perform similar validation checks, while in other cases it may happen that some quality checks are performed neither by MS nor Eurostat.

The process suffers from a series of inefficiencies: lack of coordination, but also lack of documentation, lack of formalisation of validation procedures and rules, low harmonisation of software solutions. Even when they are properly documented, validation rules are described using specific languages developed by individual process managers independently from each other, instead of using a common harmonised syntax. This situation raises several issues related to quality assessment (data completeness, accuracy, timeliness and punctuality,…) while hampering the integration of validation solutions and, as a consequence, the perspectives of a large-scale reduction of IT development and maintenance costs at European level.

These issues require the elaboration of a comprehensive solution, which calls for a portfolio of actions. At European level, a comprehensive project on data validation has been launched in the framework of the ESS Vision 2020 [1]. The scope of this document is to present an international activity, which is coordinated with the ESS action mentioned above, and is specifically addressing the issue of the lack of a standard syntax for expressing validation and editing rules.

2. **The New Validation and Transformation Language (VTL)**

Building on the SDMX [2] ISO standard for data and metadata exchange, a task-force was formed in 2013 with the purpose of elaborating a formal and standard framework for the description of logical algorithms to validate statistical data and calculate derived data.

SDMX already has, in its information model, a module for "transformations and expressions", although a specific language did not exist. To make this framework operational, a standard “language” for defining validation and transformation rules (set of operators, their syntax and semantics) is needed, together with appropriate IT formats for exchanging rules and related metadata, and web services to store and retrieve them.

The intention is to provide a language which is usable by statisticians to express logical validation rules and transformations on data, whether described as dimensional tables or as unit-level data. The assumption is that this logical formalization of validation and

---

1 Eurostat, Unit B1: Methodology and corporate architecture – Standards Team
transformation rules would be converted into specific programming languages for execution (SAS, R, Java, SQL, etc.) but would provide a “technology-neutral” expression at business level of the processing taking place, against which various implementations can be mapped. Experience with existing examples suggests that this goal is achievable.

An important point that emerged is that, besides SDMX which was the starting point, other information standards (such as GSIM [3] and DDI [4]) were also interested in such a language. However, each standard operates on its model objects and produces objects within the same model: to cope with this, the VTL language has been built upon a very basic information model, taking the common parts of GSIM, SDMX and DDI. This way, existing technical standards (SDMX, DDI, others) may adopt VTL by mapping their information models against the VTL one. Therefore, although a work-product of SDMX, the VTL language will be usable with other standards as well.

In the VTL model, both unit and dimensional data are considered as mathematical functions having independent and dependent variables and are treated in the same way. For each Unit (e.g. a person) or Group of Units of a Population (e.g. groups of persons of a certain age and civil status) identified by means of the values of the independent variables (e.g. either the “person id” or the age and the civil status), the mathematical function provides for the values of dependent variables, which are the properties to be known (e.g. revenue, expenses …). This way, the manipulation of any kind of data (unit and dimensional) is brought back to the manipulation of very simple and well-known objects, which can be easily understood and managed by users.

3. **The High-Level Characteristics of the Validation and Transformation Language (VTL)**

The task-force identified the main characteristics that the language should follow:

- **User orientation**
  - designed for the users, who should be able to define transformation and validation expressions autonomously, without IT skills and IT people intermediation;
  - intuitive and friendly (users should define and understand expressions as easily as possible);
  - oriented towards statistics, which is the main user skill; it should be possible to include operators specifically needed in the statistical process (for example operators for data validation, editing and imputation, time-series processing, …).

- **Integrated approach**
  - independent of the statistical domain of the data to be processed;
  - mapped unambiguously to the proper information model (IM); in other words, it should be able to operate on IM artefacts and to produce other IM artefacts (this is a basic property of any robust language).
  - suitable for various typologies of data of a statistical environment (for example dimensional data, survey data, registers data or transactions, micro and macro, quantitative and qualitative, …), as much as they are supported by the IM;
  - independent of the steps of the statistical process (GSBPM [5]) and usable in any one of them.

- **IT implementation independence**
  - not oriented to a specific IT implementation but allowing many different implementations (this property is particularly important for a standard language, which should not be tied to a specific IT solution and should allow different institutions to rely on different IT environments);
able to support the possible use of various IT tools in an integrated IT solution (for example, different calculation tools in different steps of statistical data processing);

make users unaware of the IT solution as much as possible;

avoid impacts on users as much as possible if the IT solution changes (for example following the adoption of another IT tool).

• **Active role for processing**
  ✓ described formally as for its grammar, to be easily parsed and processed (i.e. in Backus-Naur Form);
  ✓ able to drive the software that perform calculations, so automatically convertible in the languages of the IT tools used for calculations (once the language is defined, it might be useful to support its conversion in some widely used IT languages, for example SQL, R, XML languages …);
  ✓ able to generate validation results that can be unambiguously interpreted by software and as much as possible easily interpretable by statisticians.

• **Extensibility and customizability**
  ✓ to introduce new VTL operators according to evolution of the business needs (e.g. the operators for the validation first and those for the compilation after);
  ✓ able to include operators derived from other languages / tools (e.g. “SQL like” operators, operators for time series processing …);
  ✓ able to customize the operators for specific needs, for example of specific organizations / specific processing (note that this requirement is typically not fulfilled by the IT languages that have a fixed list of operators).

• **Proper governance**
  ✓ It implies the creation of appropriate governance rules to control the evolution of the language;
  ✓ In addition to the standard, there is the need for allowing customized parts of the language under the private governance of single institutions, which may integrate the language for their own purposes; therefore coordinated governance rules between the standard part and the customized parts should be introduced.

• **Language effectiveness**
  The effectiveness is connected to some aspect of the language features, for example:
  ✓ Historicity: possible changes of the artefacts or its sets with reference to the change of time;
  ✓ Persistency control: possibility of specifying the persistency of the intermediate results;
  ✓ Expressions chaining: possibility of expressions having also other expressions as an input parameter;
  ✓ Strictly defined (or clearly stated when undefined) behaviour for missing data, multi measures, data attributes.

4. **RESULTS**

In February 2015, after a public review period in autumn 2014, the VTL 1.0 package has been published on the web at [http://www.sdmx.org](http://www.sdmx.org). The set includes:

a) General part, highlighting the main characteristics of VTL, its core assumptions and the information model the language is based on;

b) Full library of operators ordered by category, including examples;

c) BNF notation (Backus-Naur Form) which is the technical notation used as a test bed for all the examples throughout the document.
The operators included in this 1.0 version of VTL are summarized in the diagram below.

5. CONCLUSIONS

VTL 1.0 contains a high-level definition of the general characteristics of the language and a list of operators for validation and transformation, as well as a simple information model on which the VTL can operate. VTL is usable by statisticians to express logical validation rules and transformations on data, whether described as dimensional tables or as unit-level data. The assumption is that this logical validation and transformation rules provide a “technology-neutral” expression at business level of the processing taking place, against which various implementations can be mapped.

The specifications for exchanging VTL validation rules in SDMX messages, for storing rules and for requesting validation rules from web services will be provided in a specific update to the SDMX Technical Standards on which the task-force is working on. This first implementation exercise will allow a further fine-tuning and bug-fixing of the first version, leading to a VTL 1.1 within one year. At Eurostat, VTL will be primarily implemented through SDMX. Implementation tests are already foreseen with some pilot domains.
6. REFERENCES


[3] UN/ECE, Generic Statistical Information Model 1.1 (GSIM),
http://www1.unece.org/stat/platform/download/attachments/97356610/GSIM%20Specification%201.1.docx?version=3&modificationDate=1388474361592&api=v2


[5] UN/ECE, Generic Statistical Business Process Model 5.0,
http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0