Data validation in business statistics

Preface

This chapter on data validation in the European Statistical System (ESS) is a part of the online publication European Business Statistics manual. It focuses on approaches for validating the quality of national output data for European purposes. It does not refer to corrective actions. Details on validation of the output data (or datasets) as expected by Eurostat are specified in the domain-specific compilation guides.

The aim of clarifying and streamlining validations within the statistical production chain is to reduce the burden and achieve better data quality in the ESS.

Considering that one of the main goals of European Business Statistics (EBS) is to improve data consistency between business domains, a specific challenge facing EBS is validation across business domains. Consistency checks should be carried out in most cases for similar concepts or ones that are the same. This involves first making an inventory across domains of similar concepts/ones that are same and checking differences in values collected between these concepts. If values are different, these differences should then be either justified or corrected. Special attention should be paid to values that are different only for a subset of countries. This may reflect the fact that some countries use the same concepts (definition, statistical unit, data source), whereas others use different ones – a typical case of a lack of consistency not only between domains, but also between countries.

Even in cases where it can be explained why figures may be different for a similar concept (e.g. 'ICT turnover' may be different from 'SBS turnover' due to time lag), efforts could be made to see if the domains could be further harmonised.

Consistency checks between domains could also be performed for concepts that are not similar but could be correlated.

The text in this chapter relies mainly on two of the main deliverables of the ESS.VIP Validation project:

- The ESS methodological handbook on validation (available here)
- The Business architecture for ESS validation (available here)

Both are living documents that will be reviewed on a regular basis, in particular in 2017, during the preliminary implementation phases.

Contents

1. Framework for data validation
   1.1 What is data validation?
   1.2 Validation and quality framework for official statistics
   1.3 Validation principles
   1.4 Data validation in the statistical production process
   1.5 Validation levels
   1.6 Validation life cycle
2. Target state for validation in the ESS
   2.1 Target business process
   2.2 Standards for validation in the ESS
3. Member States implementation
   3.1 Implementation options for Member States
   3.2 Validation services made available to Member States
4. See also
5. Further Eurostat information
6. External links
7. Contacts

1. Framework for data validation

1.1 What is data validation?

According to the definition in the Methodology for data validation manual, *data validation is an activity verifying whether or not a combination of values is a member of a set of acceptable combinations.*

The set of 'acceptable values' may be a set of possible values for a single field. But under this definition it may also be a set of valid value combinations for a record, column, or larger collection of data. We emphasise that the set of acceptable values does not need to be defined extensively. This broad definition of data is introduced to make data validation refer both to micro and macro (aggregated) data.

Data validation assesses the plausibility of data: a positive outcome will not guarantee that the data is correct, but a negative outcome will guarantee that the data is incorrect.

Data validation is a decisional procedure ending with the acceptance or refusal of data. The decisional procedure is generally based on rules expressing the acceptable combinations of values. Rules are applied to data. If data satisfy the rules, which means that the combination expressed by the rules is not violated, data are considered valid for the final use they are intended to. There is of course the possibility of using the complementary approach in which rules are expressed in 'negative form': in this case data are validated by verifying that predefined non-acceptable combinations of values do not occur.

Sometimes the rules used in a validation procedure are split in hard/fatal edits and soft/query edits and the not acceptable values are classified either as 'erroneous' or 'suspicious' depending on whether they fail hard edits or soft edits. Hard edits are generally rules that *must* necessarily be satisfied for logical or mathematical reasons (e.g., children cannot be older than their parents). An example of query edits taken from the UNECE glossary on statistical data editing is 'a value that, compared to historical data, seems suspiciously high' while for fatal edits is 'a geographic code for a Country province that does not exist in a table of acceptable geographic codes'. This distinction is important information for the related 'editing' phase. In addition to this information, a data validation procedure may assign a degree of failure (severity) that is important for the data editing phase and for the tuning of data validation. Taking the example previously mentioned for soft edits, the severity can be evaluated by measuring the distance of the actual values with respect to the historical one.

In case of failure of a rule, data are exported from the data validation procedure or marked respectively, and are handled by the editing staff in order to correct values to make the rules satisfied, or data are considered acceptable and the rules of the data validation are updated. The data validation process is an iterative procedure based on the tuning of rules that will converge to a set of rules that are considered the minimal set of relations that must be necessarily satisfied.

1.2 Validation and quality framework for official statistics

The formalisation of validation within the ESS needs to be considered against the backdrop of the quality standards for official statistics. The key supporting document is the European Statistics Code of Practice (available here).

The European Statistics Code of Practice is based on '15 Principles covering the institutional environment, the statistical production processes and the output of statistics. A set of indicators of good practice for each of the Principles provides a reference for reviewing the implementation of the Code. The quality criteria for European Statistics are defined in European Statistical Law.

Statistical authorities, comprising the Commission (Eurostat), National Statistical Institutes and other national authorities responsible for the development, production and dissemination of European Statistics, together with governments, ministries and the European Council, commit themselves to adhere to the Code.

The Principles of the Code of Practice together with the general quality management principles represent a common quality framework in the European Statistical System.'

Clarification of the validation checks to be performed on data produced by the ESS plays a key role in compliance with the 15 Principles of the European Statistics Code of Practice. In particular, Principles 11 to 15 below, which refer to 'statistical output', benefit greatly from clarification of the validation checks:

- Principle 11: Relevance (European Statistics meet the needs of users);
- Principle 12: Accuracy and Reliability (European Statistics accurately and reliably portray reality);
- Principle 13: Timeliness and Punctuality (European Statistics are released in a timely and punctual manner);
- Principle 14: Coherence and Comparability (European Statistics are consistent internally, over time and comparable between regions and countries; it is possible to combine and make joint use of related data from different sources); and

- Principle 15: Accessibility and Clarity (European Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance).

1.3 Validation principles

In addition to those described in the European Statistics Code of Practice, six further principles were drawn up for the validation processes (see Annex A of the Business Architecture for ESS Validation). These six principles are fully compatible with those in the European Statistics Code of Practice, although their aim is to provide guidance specifically on how to improve the validation processes. They are particularly relevant for designing the business and IT architecture for data validation.

1. THE SOONER, THE BETTER
   Validation processes must be designed to be able to correct errors as soon as possible, so that data editing can be performed at the stage where the knowledge is available to do this properly and efficiently.

2. TRUST, BUT VERIFY
   When exchanging data between organisations, data producers should be trusted to have checked the data before and data consumers should verify the data on the common rules agreed.

3. WELL-DOCUMENTED AND APPROPRIATELY COMMUNICATED VALIDATION RULES
   Validation rules must be clearly and unambiguously defined and documented in order to achieve a common understanding and implementation among the different actors involved.

4. WELL-DOCUMENTED AND APPROPRIATELY COMMUNICATED VALIDATION ERRORS
   The error messages related to the validation rules need to be clearly and unambiguously defined and documented, so that they can be communicated appropriately to ensure a common understanding on the result of the validation process.

5. COMPLY OR EXPLAIN
   Validation rules must be satisfied or reasonably well explained.

6. GOOD ENOUGH IS THE NEW PERFECT
   Validation rules should be fit-for-purpose: they should balance data consistency and accuracy requirements with timeliness and feasibility constraints.

1.4 Data validation in the statistical production process

The business processes for the production of official statistics are described in the Generic Statistical Business Process Model (GSBPM, UNECE 2013).

The schema in GSBPM shows that data validation is performed during different phases of a production process. The phases where validation is performed are as follows:

GSBPM: sub-phase 2.5

The first phase in which data validation is introduced is the 'design' phase, more specifically sub-phase 2.5, '*Design processing and analysis*'. The description in GSBPM is as follows:

'*This sub-process designs the statistical processing methodology to be applied during the 'Process' and 'Analyse' phases. This can include specification of routines for coding, editing, imputing, estimating, integrating, validating and finalizing data sets.'*

This is related to the design of a validation procedure, or more specifically, a set of validation procedures consisting of a validation plan.

GSBPM: sub-phase 4.3

The first sub-phase of GSBPM in which validation checks are performed is 4.3 '*Run collection*' (as part of the 'Collect' phase). As described in the GSBPM document, checks deal with the formal aspects of data and not the content:

'Some basic validation of the structure and integrity of the information received may take place within this sub-process, e.g. checking that files are in the right format and contain the expected fields. All validation of the content takes place in the Process phase.'

GSBPM: sub-phase 5.3

In the process phase, sub-phase 5.3 explicitly refers to validation, in fact it is called 'Review and validate'. The description given in the GSBPM document is as follows:

'*This sub-process examines data to try to identify potential problems, errors and discrepancies such as outliers, item non-response and miscoding. It can also be referred to as input data validation. It may be run iteratively, validating data against predefined edit rules, usually in a set order. It may flag data for automatic or manual inspection or editing. Reviewing and validating can apply to data from any type of source, before and after integration. Whilst validation is treated as part of the 'Process' phase, in practice, some elements of validation may occur alongside collection activities, particularly for modes such as web collection. Whilst this sub-process is concerned with detection of actual or potential errors, any correction activities that actually change the data are done in sub-process 5.4 (edit & impute)*'.

Several observations:

- The term 'input data validation' suggests an order in the production process. The term and the idea can be used in the manual.
- Validation may occur alongside collection activities;
- A distinction is made between validation and editing, and it is in the action of 'correction' that is performed in the editing sub-phase, while validation only expresses whether there is (potentially) an error or not. The relationship between validation and data editing will be discussed later on; and
- Even if an error is to be corrected in the editing sub-phase, in some cases errors may reveal a need to improve the design, build or collection phase of GSBPM.

GSBPM: sub-phase 6.2

The last sub-phase is 6.2 ('Validate outputs'):

'This sub-process is where statisticians validate the quality of the outputs produced, in accordance with a general quality framework and with expectations. This sub-process also includes activities involved with the gathering of intelligence, with the cumulative effect of building up a body of knowledge about a specific statistical domain. This knowledge is then applied to the current collection, in the current environment, to identify any divergence from expectations and to allow informed analyses. Validation activities can include:

* checking that the population coverage and response rates are as required;*

✓ *comparing the statistics with previous cycles (if applicable);*

* checking that the associated metadata and paradata (process metadata) are present and in line with expectations;*

✓ *confronting the statistics against other relevant data (both internal and external);*

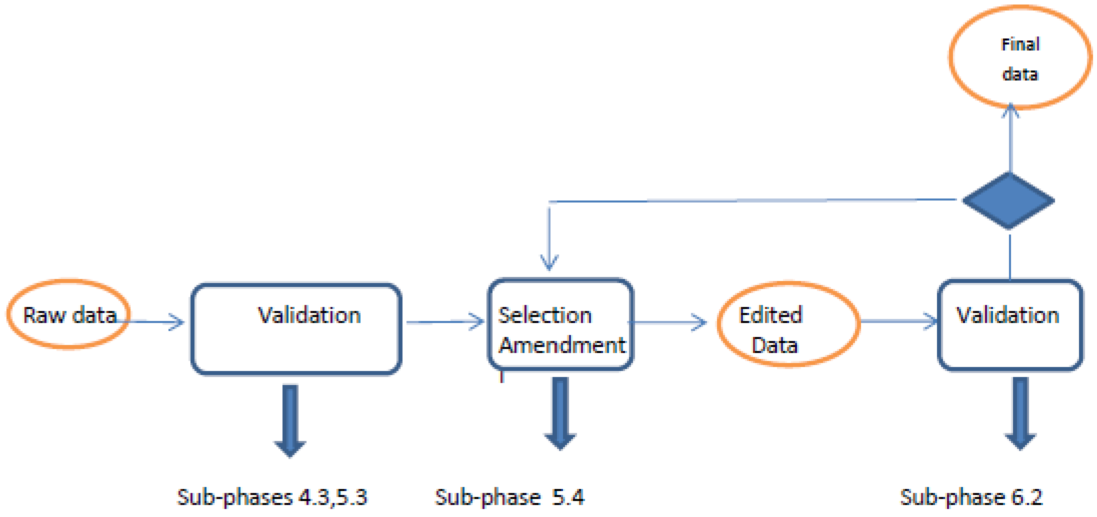
✓ *investigating inconsistencies in the statistics;*

✓ *performing macro editing;*

✓ *validating the statistics against expectations and domain intelligence'.*

The checks that are not usually considered part of a 'data validation' procedure (i.e. the first and the third item where emphasis is not on data) are marked with ''.

Figure 1: Flowchart describing the different GSBPM validation phases linked with statistical data editing



1.5 Validation levels

Examining the practical implementation of the validation process means looking at it from a business perspective. In doing so, the focus is on the validation activities.

The amount and accessibility of information needed and the phases of the validation process are important for determining the validation levels. This approach is particularly useful when classifying and designing validation activities within an organisation.

Validations could be divided into structural validations and content validations.

- Structural validations are linked to the definition of the data structure. In the SDMX (Statistical Data and Metadata eXchange) context,[1] this also includes the definition of the code lists and constraints related to the use of specific codes. Structural validations refer here to validation level 0 and a part of validation level 1 described below.

- Content validations are linked to levels 1 to 5 described below. They also rely on a clear definition of the data structure.

Validation level 0: consistency with the expected IT structural requirements.
Check e.g. that:

- The file has the expected number of columns (agreed format of the file);
- The column has the expected format (i.e., alphanumeric, numeric etc.).

Validation level 1: consistency within the dataset.
Check e.g. that:

- The content of the third column is one of the codes from the ‘Sex’ dictionary;
- The content of the first column (reporting country) is consistent with the data sender;
- Total inhabitants = male inhabitants + female inhabitants.

Validation level 2: consistency with other datasets within the same domain and data source.
Check e.g. that:

- New data referring to a new time period is not an outlier (does not vary by more than 10 % compared to data from the previous time period);
- Annual data is consistent with data from the corresponding quarterly datasets.

Validation level 3: consistency within the same domain between different data sources (mirror checks).
Check e.g. that the export declared by country A to country B is the same as the import declared by country B from country A.

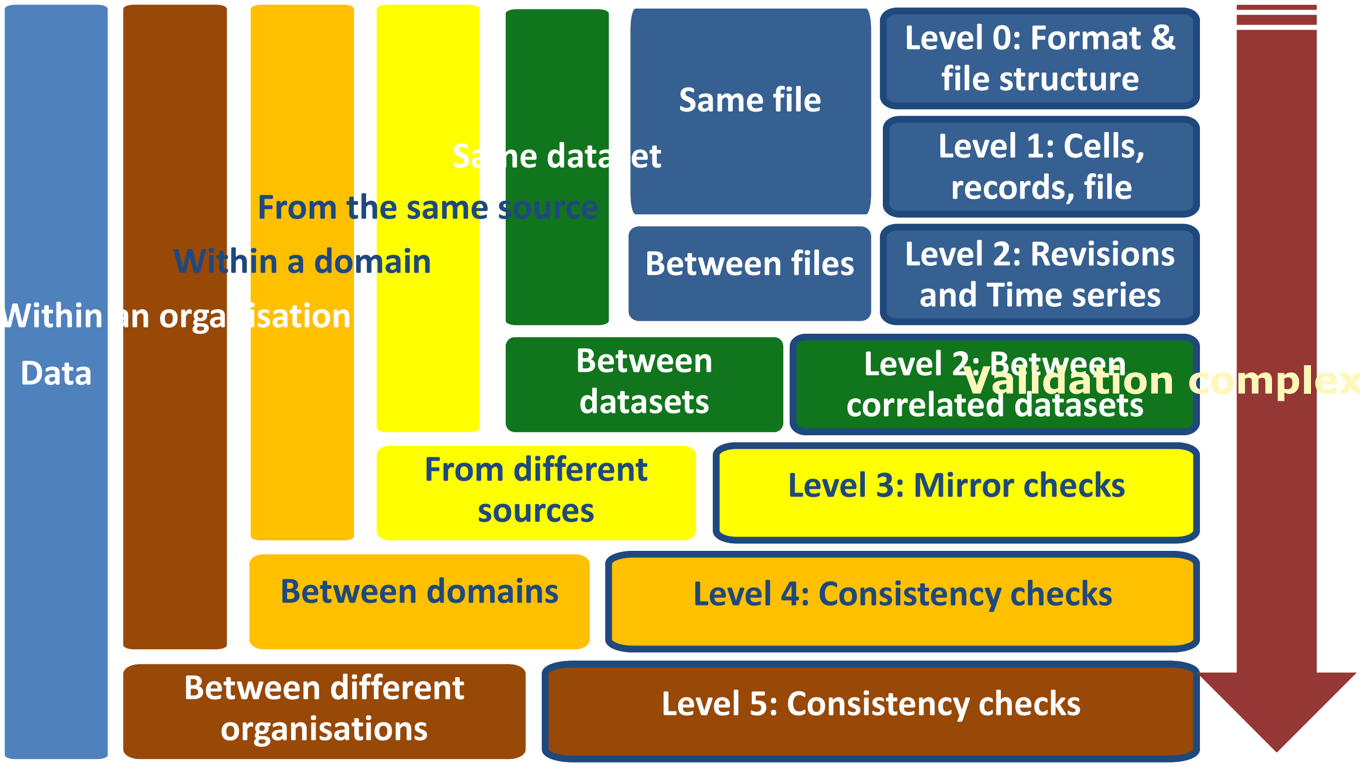
Validation level 4: consistency between separate domains available in the same organisation.
Check e.g. that the number of enterprises and employees in SBS and Business demography are consistent for the same time period.

Validation level 5: consistency with data available in other organisations.

Check e.g. that country data in the ESS are consistent with the data available in the World Trade Organisation, International Labour Organisation, World Bank etc.

[1] See section on SDMX in chapter 2.2.

Figure 2: Graphical representation of validation levels


Same file
Level 0: Format & file structure
Level 1: Cells, records, file
Same dataset
From the same source
Within a domain
Within an organisation
Data
Between files
Level 2: Revisions and Time series
Between datasets
Level 2: Between correlated datasets
Validation complexity
From different sources
Level 3: Mirror checks
Between domains
Level 4: Consistency checks
Between different organisations
Level 5: Consistency checks

1.6 Validation life cycle

To improve the performance of a statistical production process by managing and optimising the data validation process, a description of the data validation process life cycle would be helpful.

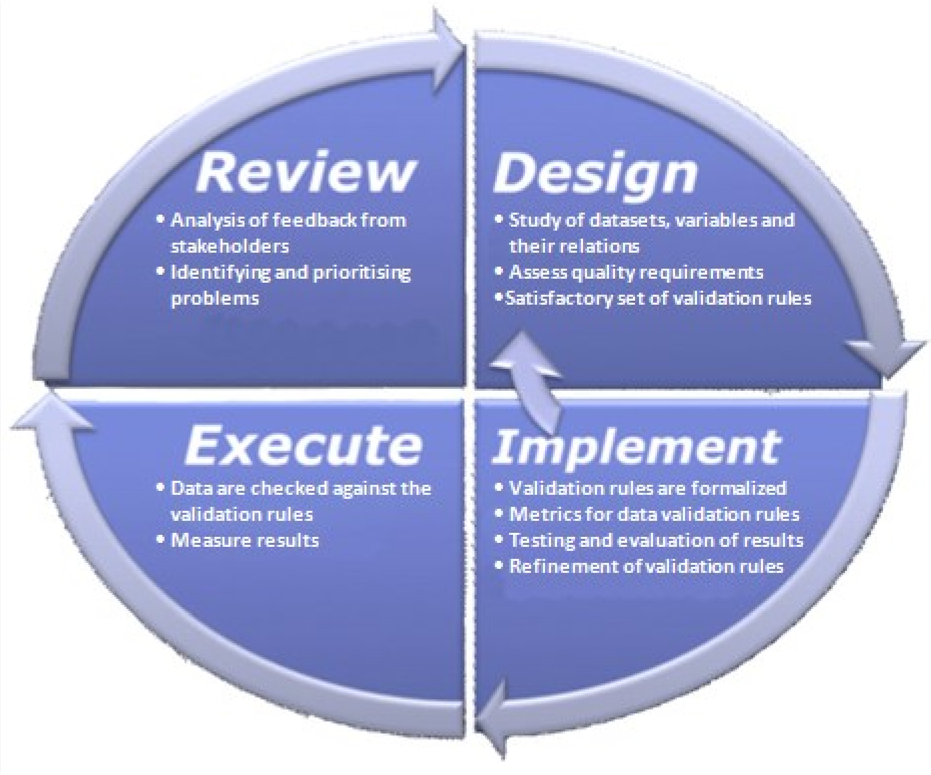
First, the process is both dynamic and complex. Adapting validation rules may affect not only the scope of one dataset or one statistical domain, but also that of all statistical domains. For instance, when optimising the effectiveness and efficiency of the validation rules, their assessment from last time, relationships with indicators etc. should be taken into account. Second, the process should be viewed as an integral part of the whole statistical information production process.

The data validation life cycle involves the activities directly linked to each statistical domain for the definition and execution of data validation. This cycle starts by **designing** the data validation process for the statistical domain or inter-statistical domain, with an overall study of the datasets, variables and their relationships to find a list of suitable and effective validation rules. In the **implementation** phase, these validation rules are described in common syntax, formalised, tested and refined, discussed and evaluated by stakeholders. During the **execution** phase, data are checked against the rules; with validation results measured and quantified. These outputs are **reviewed** to improve the list of validation rules.

The data validation process is an integral part of the whole statistical information production process. Validation tasks and controls are performed by several stakeholders with a wide range of responsibilities. The data validation process life cycle should provide a clear and coherent allocation of actions and responsibilities to ensure the highest level of performance, while reducing the number of possible errors. However, it may be difficult to allocate responsibilities to each phase of the data validation life cycle due to the complexity of the data validation procedure and because this is closely related to the specific structure of the organisation.

Designing validation rules and rule sets for a dataset involves distributing validation tasks in the statistical production chain to be proposed to the decision-making structures. This distribution of responsibilities should be designed based on the principle of *'the sooner the better'* as it is commonly agreed that the cost of fixing data errors in terms of resources, time and quality is lower the closer it is to the data source.

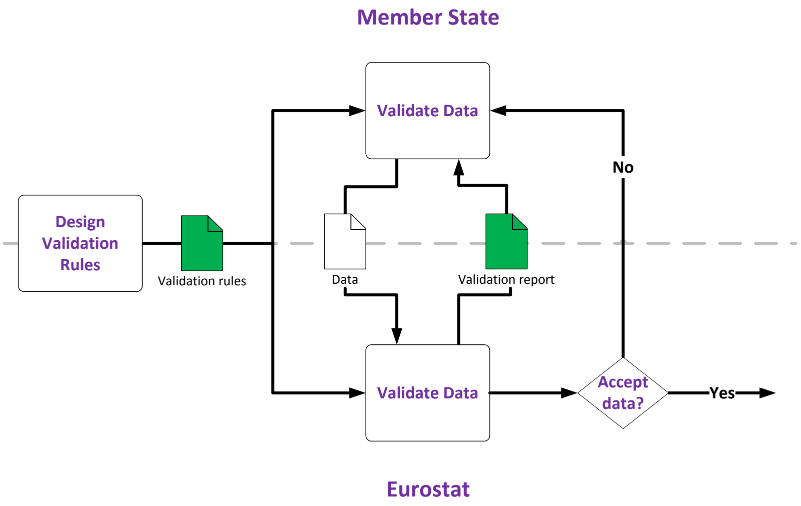
Figure 3: Data validation process life cycle


Review
• Analysis of feedback from stakeholders
• Identifying and prioritising problems
Design
• Study of datasets, variables and their relations
• Assess quality requirements
•Satisfactory set of validation rules
Execute
• Data are checked against the validation rules
• Measure results
Implement
• Validation rules are formalized
• Metrics for data validation rules
• Testing and evaluation of results
• Refinement of validation rules

2. Target state for validation in the ESS

2.1 Target business process

Figure 4: Target business process (to-be state) for validation in the ESS



In the target business process, validation rules are jointly designed and agreed upon at the level of each statistical domain's working group. The resulting validation rules are documented using common cross-domain standards, with clear validation responsibilities assigned to the different groups participating in the production process of European statistics.

The use of common standards for validation rules and validation reports, combined with the common ESS guidelines for IT services being developed by the ESS Vision 2020 SERV project, will enable shareable ESS services to be created for validation. Member States will be able to use them on a voluntary basis to validate the data to be sent to Eurostat.

2.2 Standards for validation in the ESS

The prerequisite for building a validation framework in a specific domain is to express the structure of the data and the format in a standardised and machine-readable way. The SDMX standard provides this description for a growing number of domains.

Validation rules then have to be described in a non-ambiguous standard language that can be understood by both humans and computers. This is the purpose of VTL (Validation and Transformation Language), which is being developed under the umbrella of the SDMX Technical Working Group.

Finally, ESS members are invited to use or develop whenever possible any relevant shared validation services compatible with the Common Statistical Production Architecture (CSPA) standard.

There are also plans to develop a standard for validation reports.

SDMX

The first step in the current ESS validation process is to define the structure, standard code lists and format of the data files to be sent to Eurostat. This business function is important for validation as it implicitly provides an initial set of validation rules related to the expected structure of the data file. This first step is usually conducted jointly by Eurostat and the Member States by means of consultations in each specific domain's working group. The main output is a document describing the expected data structure. In recent years, SDMX has played an increasing role in standardising this step. Around 40 % of European statistical production processes now use the SDMX formalism to describe their data structures. The emergence of SDMX has enabled shared services to be created in support of this business function (e.g. the Euro-SDMX Registry and the SDMX Global Registry).

VTL

The Validation and Transformation language (VTL) has been designed mainly for non-IT people and is geared to the statistical world.

VTL is a standard language for defining validation and transformation rules (set of operators, their syntax and semantics) for any kind of statistical data. VTL builds on the SDMX information model, but it can also be used with any kind of structured data and data typology (microdata, aggregated data, registers, qualitative, quantitative).

The logical formalisation of VTL validation and transformation rules allows several implementations using specific programming languages for execution (R, SAS, Java, SQL etc.).

The specifications for exchanging VTL validation rules in SDMX messages, storing rules and requesting validation rules from web services will be provided in a specific update to the SDMX Technical Standards, which the SDMX Technical Working Group is working on.

The VTL 2.0 specifications have been endorsed by the SDMX community in March 2018. They will form the basis for ESS shared tools/service developments and will be used as a common language for documenting ESS validation rules.

CSPA services

The Common Statistical Production Architecture (CSPA) is a reference architecture for the statistical industry. It has been developed and peer reviewed by the international statistical community as a key standard of the ModernStats initiative (Modernisation of official statistics) set up by HLG-MOS (High-Level Group for the Modernisation of Official Statistics) under UNECE-CES (Conference of European Statisticians).

CSPA:

- covers statistical production across the processes defined by the GSBPM;
- provides a practical link between conceptual standards (the Generic Statistical Information Model (GSIM) and the Generic Statistical Business Production Model (GSBPM)) and statistical production;
- includes application architecture and associated principles for the delivery of statistical services;

- does not prescribe technological environments of statistical organisations.

CSPA-compatible services aim to allow integration in a Service Oriented Architecture (SOA) and therefore support the reuse and sharing of software components in the international statistical community.

Eurostat and the ESS are currently working on the development of CSPA-compatible services for data validation.

Eurostat maintains a catalogue of CSPA services available in the ESS (accessible from the UNECE CSPA Global Artefacts Catalogue here).

3. Member State implementation

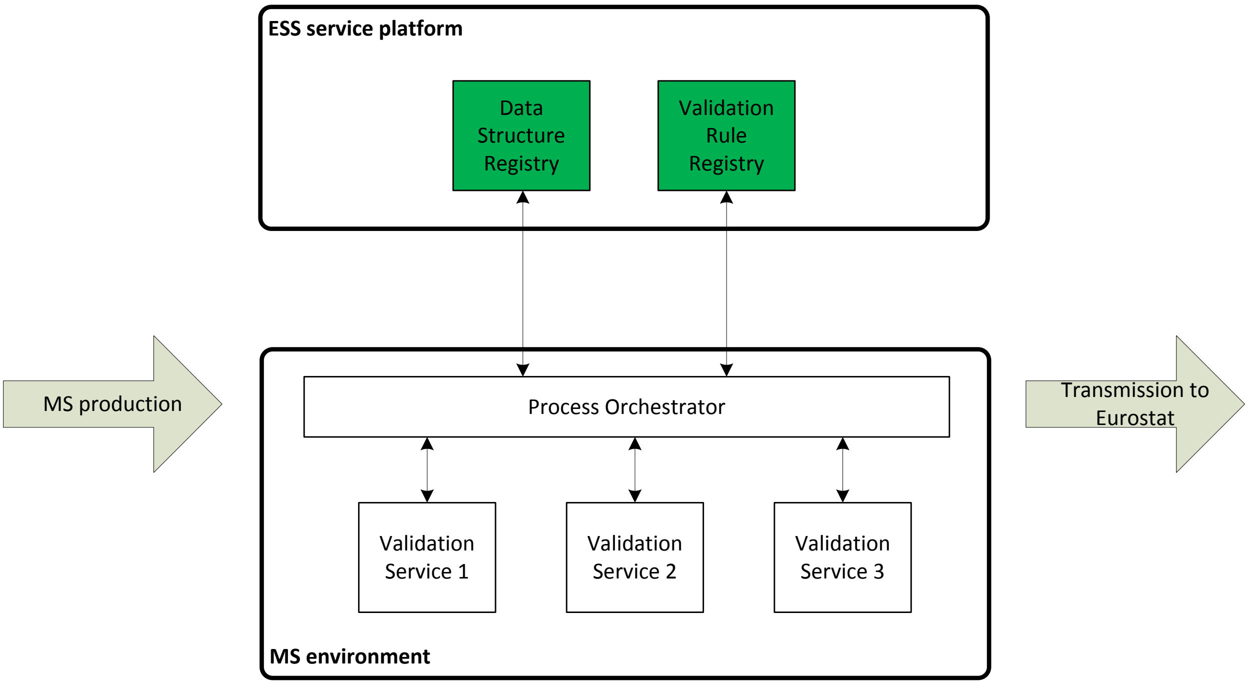
3.1 Implementation options for Member States

As long as the validation rules that have been jointly agreed upon are applied, each Member State should be able to freely choose, for each statistical production process, the extent to which it wants to benefit from the availability of reusable ESS validation services. There are therefore three basic scenarios in which Member States could implement validation rules in the target to-be state.

Scenario 1: Autonomous validation services

In this scenario, Member States would use their own autonomous services to implement the validation rules before transmitting data to Eurostat. However, these services would use the data structures and validation rules jointly agreed upon, which would be stored in centrally hosted registries. Translating these validation rules into autonomous validation services would be the responsibility of each Member State.

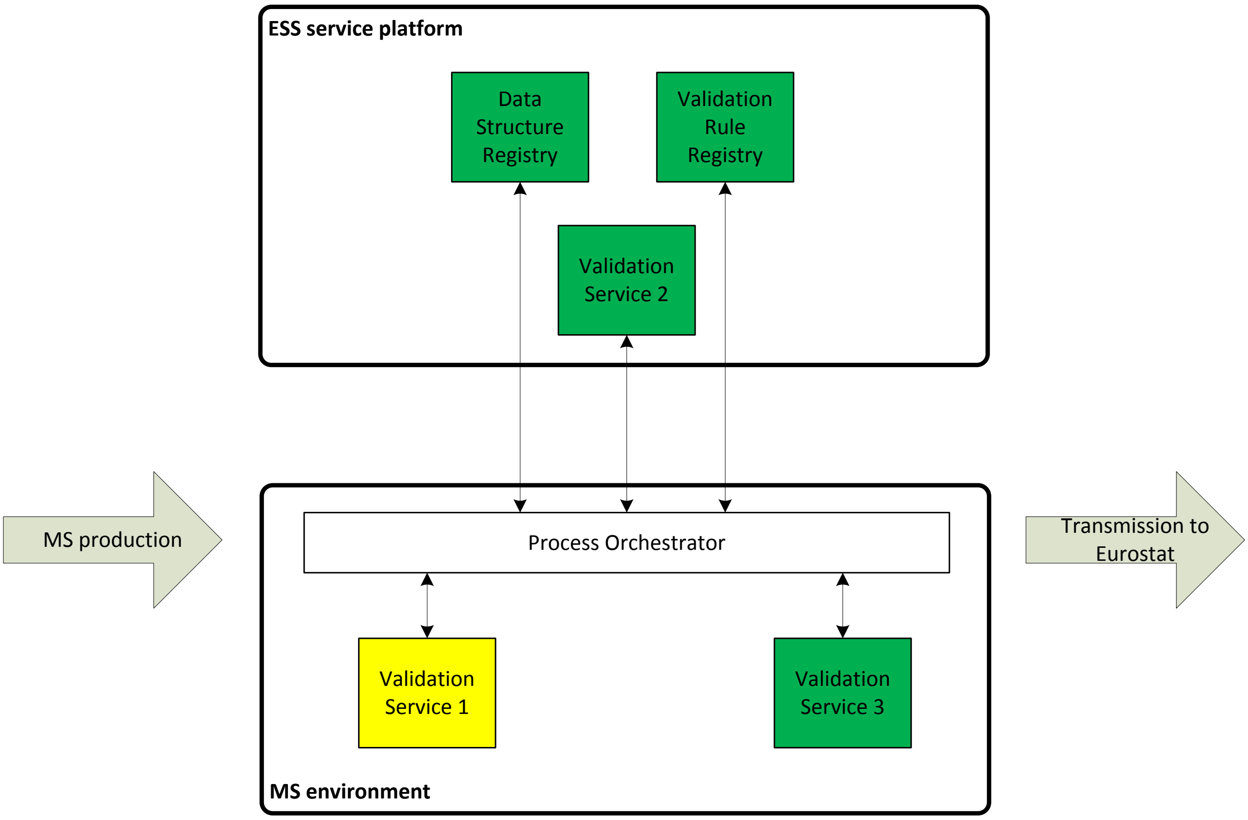
Figure 5: Scenario 1 — Member States implement the agreed validation rules in their own validation systems



Scenario 2: Replicated/shared validation services

In this scenario, in addition to the shared registries for data structures and validation rules, Member States would use certain replicated and/or shared validation services in their validation process. They would be free to select a combination of autonomous/interoperable/replicated and shared services they find most suitable. They would also still be responsible for managing the different services used in the validation process. This scenario can be likened to the Software as a Service (SaaS) model in cloud computing.

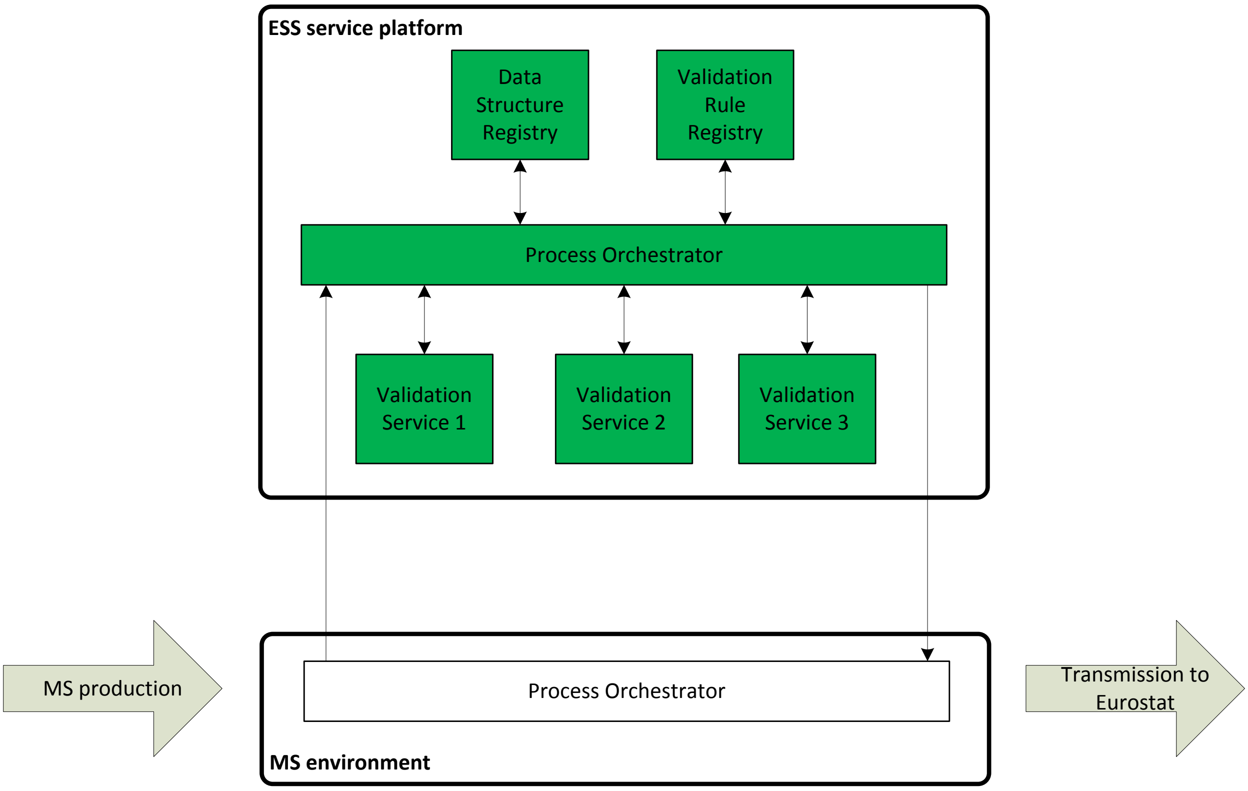
Figure 6: Scenario 2 — Member States use common ESS services in their validation process



Scenario 3: Shared validation process

In this scenario, Member States would delegate the validation of their data to a shared validation process that is predefined centrally. This process would manage the various services needed and would provide the Member State with a comprehensive validation report. This scenario can be likened to the Business Process as a Service (BPaaS) model in cloud computing.

Figure 7: Scenario 3 — Member States submit their data to a shared validation process (predefined centrally)



The scenarios above represent rather idealised situations. In real-life situations, it is likely that Member States will create hybrid scenarios that incorporate elements from two or more scenarios. Each Member State would be free to mix and match the three scenarios as it sees fit.

3.2 Validation services available to Member States

The validation services are IT tools made available by Eurostat to Member States. They are CSPA-compatible. They can either be integrated into the production system of the Member States, called from the production system or used as central services to which data are submitted for validation (variants of scenarios 2 and 3 described above).

STRUVAL (Structural validation)

The STRUVAL service performs structural validation of statistical data files based on structural information in accordance with the SDMX information model for a given data flow. It ensures that a data file respects the structure and coding of the DSD (Data Structure Definition) and the constraints defined for the respective data flow.

It also checks that the physical format used to transmit the data is compatible with the expected SDMX format (SDMX-ML, SDMX-CSV etc.).

STRUVAL started to be used in summer 2016 for NAPS-S (National Account Production System — Services).

CONVAL (Content validation)

The CONVAL service performs the validation of the content of statistical datasets based on validation rules and constraints formulated by the statistical domain managers responsible for the respective business processes and datasets. This generic validation service can be used via a graphical user interface or by connecting to a process manager layer of the service architecture that executes a configured workflow. The service is a key component of the data validation process performed by Eurostat and the ESS.

The service provides and performs the complete range of validation operations employed by statistical production. It carries out basic logical checks and content checks, intra- and inter-file data plausibility and consistency checks, and cross-domain, source-based checks.

In addition, the service informs stakeholders of the validity and consistency of datasets. It produces a validation report as output, which is distributed to stakeholders of the validation service. The report contains errors that were detected separately and classified to support effective business response.

An initial version of CONVAL (version 1.0) was released in 2017. It is based on EDIT, a validation tool used by more than 20 statistical domains in the ESS. The first version of CONVAL compatible with VTL 2.0 is expected in 2018 (CONVAL version 2.0).

Validation Rule Manager

The Validation Rule Manager will enable statisticians to express validation rules in a user-friendly way. For the most common types of validation rules in the ESS, users will specify some parameters and the VTL script will be generated automatically, allowing users to create and maintain these rules without knowledge of VTL. The system will maintain a registry of validation rules with their associated metadata.

A typical use case for the Validation Rule Manager is as follows:

- Eurostat set ups and maintains the validation rules;
- The shared validation services use the validation rules;
- Specific validation services developed by Member States also use the validation rules; and

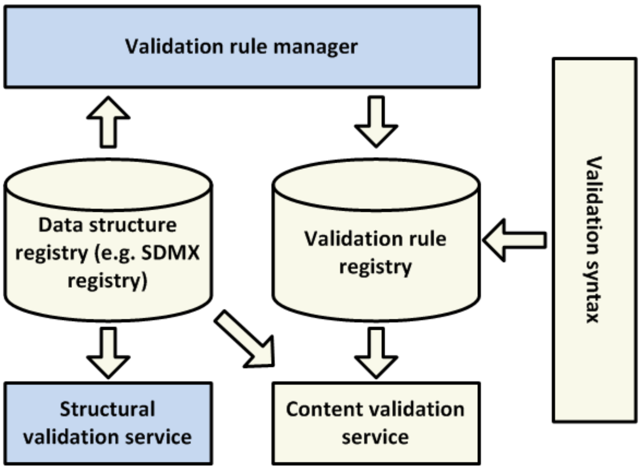
- Users in Member States could also directly access the validation rules applicable to their domains/datasets.

The Validation Rule Manager and the Validation Rule Registry are expected to be available in 2019.

SDMX tools and services

SDMX tools and services such as the SDMX registry, SDMX RI or SDMX converters may also be used by Member States together with validation tools and services in their business architecture.

Figure 8: Validation in the ESS — IT building blocks


Validation rule manager
Data structure registry (e.g. SDMX registry)
Validation rule registry
Validation syntax
Structural validation service
Content validation service

4. See also

- Overview of methodologies of European business statistics: EBS manual
- Legal provisions related to Data validation can be found in the following overview

Domain-specific validation rules (incl. Data Structure Definitions)
<placeholder for putting links to validation rules applicable to the business domains>

- STS <an example is provided in the annex>
- ICT
- SBS
- …

5. Further Eurostat information

- ESS manual on methodology for data validation
- Business Architecture for ESS Validation
- European Statistics Code of Practice

6. External links

- Statistical Data and Metadata eXchange (SDMX) website
- Validation and Transformation Language (VTL) on the SDMX website
- UNECE – Generic Statistical Business Process Model (GSBPM)
- UNECE – Common Statistical Production Architecture (CSPA)
- UNECE – CSPA Global Artefacts Catalogue (incl. link to the CSPA service catalogue in the ESS)
- UNECE – High-Level Group for the Modernisation of Official Statistics (Modernstats)

7. Contacts

ESTAT-VALIDATION@ec.europa.eu

Annex: Example of a validation rule for STS expressed in VTL 1.1

Rule: STS_1C_CEC_2 (consistency between most recent observation period in data file and the envelope)

Rule type: Consistency between envelope and content

Link with reference document (SDMX for STS — Appendix 3 — data validation rules**):** Inspired from part of rule 1

Description rule:

The most recent observation period in the data file must correspond to the observation period of the envelope.
i.e. the most recent combination of year and quarter or year and month in the dimension « TIME_PERIOD » of the data file must correspond to the year/period identified in the eDAMIS flow.

Refers to: All STS datasets

Data structure: STSALL

FREQ; REF_AREA; ADJUSTMENT; INDICATOR; ACTIVITY;BASE_YEAR; **TIME_PERIOD;** OBS_VALUE; OBS_STATUS; CONF_STATUS; UNIT_MULT; UNIT; DECIMALS; TRANSFORMATION; PRE_BREAK_VALUE; TIME_FORMAT; COMMENT_DSET; COMMENT_OBS; EMBARGO_TIME; COMMENT_TS

Severity: Error

VTL:

Parameters:

ds_sts the dataset to be validated

end_date the end date (e.g. 2015- Q4 if years and quarters, or 2015-12 if years and months), derived from the envelope in the eDamis flow

Approach:

Check if most recent date (maximum value) in the TIME_PERIOD dimension corresponds to the end date.

Returns:

Empty dataset if correct — otherwise the most recent (incorrect) time period along with the corresponding error level and error code.

VTL code:

ds_result_2:= check(max (ds_sts.TIME_PERIOD) = end_date, errorcode (‘The most recent date does not correspond to the time period specified in the eDamis flow’), errorlevel (‘E’));

Example:

For a data file identified in eDAMIS with the following envelope « STSCONS_PERM_M_HU_2015_0010_Vnnnn.xxx» (Building permits, number of dwellings or square metres of useful floor area — Monthly data for Hungary — 2015 — October):

Good data file

because the most recent TIME_PERIOD in the records is 2015-10:

For example in the record below:

M;HU;N;PNUM;F_CC1 1_X_CC1 13;ABS 0;**2015-10**;1011;A;F;;;;;0;PN;;;P1M ;

Bad data file

due to at least the record below:

M;HU;N;CSTM;F_CC1 1_X_CC1 13;ABS 0;**2015-11**;1011;A;F;;;;;0;PN;;;P1M ;

=> 2015-11 (Nov 2015) is more recent than the year/month of the envelope.

Bad data file

=> the most recent TIME_PERIOD in the data file is 2014-12 (Dec 2014), which is older than the year/month of the envelope (2015-10)

For example in the record below:

M;HU;N;PSQM;F_CC1 12;ABS 0;**2014-12**;56223;A;F;;;;;0;PN;;;P1M ;