Working Paper 13.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

The modernisation of validation in the ESS – a multidimensional approach

Prepared by Luca Gramaglia and Vincent Tronet, Eurostat

I. Introduction

One of the defining characteristics of the production of European statistics is the fact that the production process is distributed across several organisations. Data collection and a first round of processing are under the responsibility of ESS Member States. The data are then transmitted to Eurostat, where the data are processed further and finally disseminated at European level.

Data validation activities occur at several points in this statistical production chain. However, one key step in the ESS statistical production is the validation of the data sent to Eurostat by Member States (see figure 1). This is the step that ensures that the data coming from different national authorities abide by common consistency and coherence requirements and is thus essential in turning national statistics into European statistics.

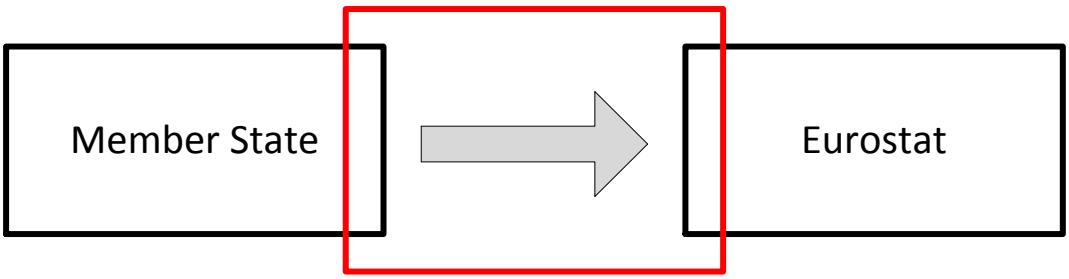


Figure 1: An important step in the production of European statistics is the validation of the data as they cross the boundary between Member States and Eurostat

In May 2014, the European Statistical System adopted the ESS Vision 2020, a common European response to the challenges facing official statistics. As the ESS Vision 2020 puts a strong emphasis on investing in quality as a defining asset of official statistics, the European Statistical System launched an ambitious programme to improve the way data sent to Eurostat are validated. This paper presents the multi-pronged approach chosen by the ESS to achieve its medium-term objectives for validation. The paper also presents a concrete example of how this approach is being implemented in the framework of the re-engineering of the National Accounts production processes in Eurostat.

II. Validation in the ESS: Challenges and objectives

The validation of the data sent by Member States to Eurostat is a joint effort involving both Eurostat and national data providers (i.e. NSIs or other national administrations). The overall quality of the ESS data

validation process is therefore heavily dependent on the quality and depth of the collaboration between Eurostat and national data providers.

While they vary considerably across different domains, current validation practices exhibit shortcomings which could be corrected through strengthened collaboration. The main such shortcomings are listed below:

- In several domains, the lack of a clear repartition of validation responsibilities among the different partners involved in the production process leads to double-work in the ESS and to the risk of "validation gaps", i.e. to cases where essential validation procedures are not carried out by any of the actors.

- The lack of shared and easily accessible documentation on validation procedures can lead to time-consuming misunderstandings between Eurostat and ESS data providers when data validation problems arise (this phenomenon has been dubbed "validation ping-pong"). It can also lead to difficulties in assessing whether the quality assurance mechanisms applied to data sent to Eurostat are "fit-for-purpose".

- The lack of common standards for validation solutions leads to a duplication of IT development and integration costs in the ESS. Moreover, the ESS is currently incurring high opportunity costs by not exploiting the general trend in the IT world towards Service-Oriented Architecture (SOA) and its potential benefits in terms of reuse and sharing of software components.

In order to respond to the flaws outlined above, the European Statistical System agreed on improving the data validation process by pursuing the two following medium-term goals:

- **Objective 1:** Ensure the transparency of the validation procedures applied to the data sent to Eurostat by the ESS Member States through a common validation policy focusing on the attribution of validation responsibilities among the different actors in the production process of European statistics.

- **Objective 2:** Improve the interoperability between Eurostat and Member States through the sharing and re-use of validation services across the ESS on a voluntary basis.

III. A multidimensional approach

Achieving the medium-term goals mentioned above requires a combination of investment along five different dimensions: methodology, processes/governance, information standards, IT and human resources.

A. Methodology

A precondition for a transparent validation process and for the identification and design of interoperable IT services is the availability of a common language for validation in the ESS. At the outset of our modernisation efforts, it became clear that the terminology used to identify concepts related to data validation varied widely across Member States and across statistical domains. Moreover, while a large literature describing specific data validation or data editing methods is available, few internationally recognized references providing a comprehensive conceptualization of the data validation process itself exist.

The ESS therefore created a methodological handbook on data validation [1]. By providing common definitions, common classifications and common metrics for data validation, the handbook gives a reference framework on which all further developments can be based.

B. Processes / Governance

A preliminary analysis of how data validation is performed in different statistical domains showed that, while the business needs are in most cases very similar, the approaches to data validation exhibit a high degree of diversity. This diversity was an obstacle to achieving the two medium-term objectives for validation in the ESS, as it often led to insufficient traceability of the validation process and to a fragmented IT landscape. The definition of a common, cross-domain approach was therefore needed.

To this end, building on the common methodological framework mentioned in the previous section, a common Business and IT architecture for validation was created [2]. The Business and IT architecture identifies a target validation business process and a series of validation principles to be implemented in all domains of European statistics (the list of validation principles can be found in Annex I). The target validation business process foresees that Eurostat and Member States will jointly define, at domain-specific Working Group level, the validation rules which must be applied to the data. Working Groups will also assign a severity level to each validation rule and determine which organisation is responsible for applying it. The Business and IT architecture also proposes a standard process for determining whether the data can be deemed acceptable.

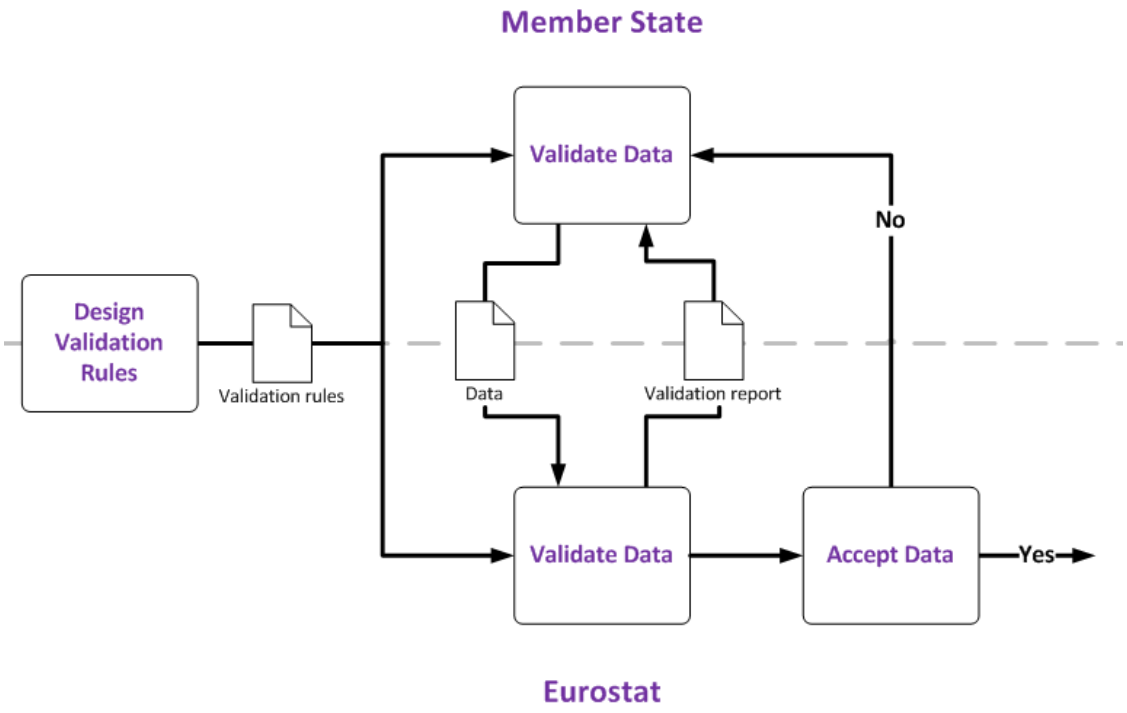


Figure 2 High-level view of the main business functions and data flows of the Business and IT archietcure for validation in the ESS

The Business and IT architecture for validation in the ESS constitutes the lynchpin for all other developments related to validation. It identifies the main data flows that will be needed to sustain an efficient validation process in the ESS, thus defining requirements for new information standards. It also determines which interoperable IT building blocks will be needed to implement the to-be state and outlines various scenarios for how Member States could, on an optional basis, profit from reusable validation services to validate their data prior to transmission to Eurostat.

C. Information Standards

In the target validation business process, Member States will exchange several kinds of information objects: data, validation rules and validation reports. A standardization of these objects is therefore

essential in order to guarantee the transparency of the validation process and to create interoperable IT services capable of using and producing them.

Eurostat therefore collaborated with several other organisations at national and international level to create a standard language to express validation rules. The result of these efforts is the Validation and Transformation Language (VTL). VTL was designed mainly for non-IT people and allows validation and transformation rules to be expressed at a logical level, while specific programming languages can be used for execution, such as R, SAS, Java or SQL [3]. VTL builds on the SDMX information model but it can be used with any kind of structured data and data typology (micro data, aggregated data, registers, qualitative, quantitative, etc.). Version 1.1 of VTL is scheduled to be released in the spring of 2017.

At the same time, over the course of 2017 a group of ESS national statistical institutes will work on creating a standard for validation reports.

D. Information Technology

The target validation business process will need to be supported by relevant IT services. In order to maximize the potential reuse by ESS Member States, these services are being developed in line with the guidelines provided by the Common Statistical production Architecture (CSPA).

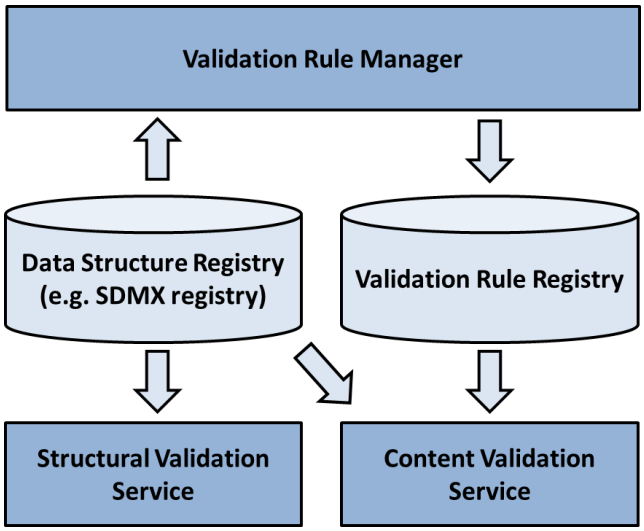


Figure 3 Overview of the validation services being developed

The figure above provides a high-level overview of the currently envisaged ESS validation services. The three main services (represented by the three rectangles) are the following:

- A structural validation service, which checks that the data comply with the appropriate data structure (e.g. correct format, correct codes, etc…). A data structure registry would supply the structural validation service with the required information on the expected data structure. A first version of the structural validation service was released in 2016.

- A content validation service, which goes beyond data structures and checks the consistency and plausibility of the data themselves (e.g. aggregation checks, detection of suspicious values etc…). A validation rule registry would supply the content validation service with the required validation rules expressed in VTL to be applied to a specific dataset. A prototype for the content validation service is being tested in a limited number of statistical domains. A first production version is expected to be released in early 2018.

A Validation Rule Manager, which would allow authorised users to view, create, modify and manage VTL validation rules stored in the validation rule registry. A first version of the Validation Rule Manger is expected to be released in 2018.

E. Human Resources

The correct implementation of the target validation business process requires communication and training on all the different aspects touched upon by the modernisation effort.

Accordingly, as more and more pieces of the target validation business process become available, communication and training actions will be stepped up. Several regional conferences on validation will take place over the course of 2017, while the first-ever ESS-wide training course on validation will take place at the end of 2017.

IV. Early results from a concrete use case: validation in the National Accounts domain

In recent years, Eurostat has invested heavily in the evolution of the National Accounts production system from classical monolithic stovepipe systems that are highly optimised for a specific statistical domain or product into a flexible service-based production architecture [4]. Table 1 offers an overview of the expected evolution of the National Account production systems in Eurostat.

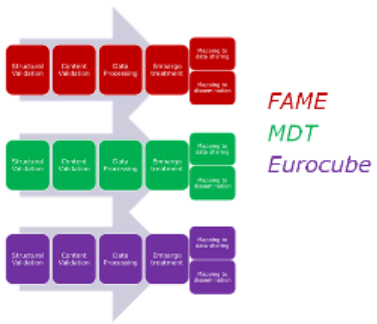
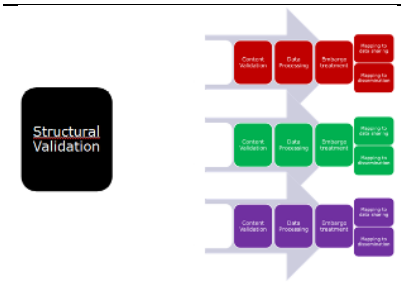
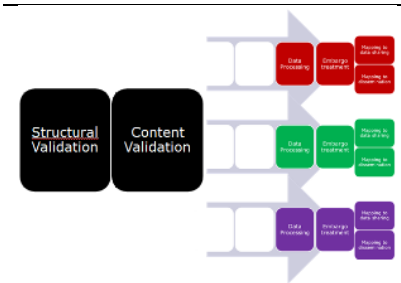
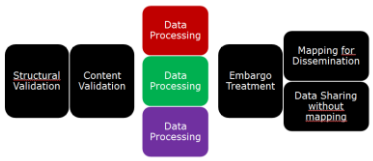
	**Initial situation**: all business process steps are reflected within the respective sub-systems.
	**Phase 1**: structural validation is performed on a flexible architecture before data enters into the sub-system. The structural validation module of the sub-system is disabled.
	**Phase 2**: content validation is added to the flexible architecture and also performed before data enters into the sub-system. Further validation modules of the sub-system are disabled.
	**Future architecture**: though not in the scope of the NAPS-S project, the next logical step would be to identify additional statistical services that can be exposed through the flexible architecture. This will require a more extensive analysis of the underlying business processes in order to identify service candidates that have the potential to be generalised.

Table 1 Overview of the implementation steps towards an SOA-based production system for national Accounts [4]. Phase 1 and 2 concentrate on implementing the validation architecture described in this paper.

As part of this endeavour, the validation process for the data sent by Member States to Eurostat has moved towards the target business process defined in the Business and IT architecture for validation in the ESS. The validation processes for the different National Accounts sub-domains were harmonised and reusable, CSPA-compliant IT services for validation were gradually introduced as they became available: a structural validation service was tested and put in production in 2016, while the same will be done for a content validation service in 2017. The National Account domain has thus been an early adopter of the new approach to validation described in this paper.

A first analysis of the outcomes of these modernisation efforts has shown that, at the end of 2016, there was a marked reduction in the number of data file transmissions from Member States to Eurostat compared to previous years. This suggests that the implementation of the target validation business process, though still partial, has already brought benefits in terms of the reduction of "validation ping-pong" between Member States and Eurostat. While additional data points will be needed to better quantify the benefits of this approach, these early results are certainly encouraging.

V. Conclusions and avenues for further work

This paper outlined the strategy pursued by the ESS to modernise the way data sent to Eurostat by Member States is validated. The multi-pronged approach proposed, while somewhat laborious to set up at first, provides in our opinion the most solid basis for long-term success. The early results of its application in the National Accounts domain give a first glimpse of its potential benefits.

The main scope of the work performed thus far has been relatively narrow: the validation of the data sent by Member State to Eurostat represents a subset of the validation activities which are carried out in order to produce European statistics. However, many of the deliverables described in this paper are generic enough to be applied to validation activities beyond the original scope. The extent to which these deliverables can be leveraged to modernise validation processes at national level is an area of active research. The ValiDat Integration ESSnet project, which was launched in early 2017 and involves six ESS Member States, is expected to contribute to answering this question.

Bibliography

[1] ESSnet ValiDat Foundation, “Methodology for data validation,” 2016.

[2] Eurostat, “Business architecture for ESS validation,” 2016.

[3] SDMX Technical Working Group, [Online]. Available: https://sdmx.org/?page_id=5096. [Accessed 03 03 03].

[4] M. V. Daniel Suranyi, “Re-engineering and re-designing statistical production - Case Study: Modernisation of National Accounts,” in *European Conference on Quality in Official Statistics*, 2016.

Annex: ESS Validation principles

The sooner, the better

a. **Statement**

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.

b. **Rationale**

This principle is at the core of any statistical validation process. There may be many reasons underlying a validation error. Finding the cause and fixing it might well include investigating the correctness of data, software, methodologies or statistical processes as a whole. This can only be done by people sufficiently familiar with the statistical domain and the way the data was produced. Hence, the sooner errors are detected in a statistical production chain, the easier and more efficient it is to correct them.

c. **Implications**

For the ESS this means that validation of national data should take place at the NSI's who have the sole responsibility for the correctness of the national data. The NSI can only do so if the validation rules are well-defined and understood (see principle 3). If national data appears to be violating validation rules after data exchange, Eurostat should inform the NSI so that correction can be done at the right place. In cases where validation errors arise from rules involving multiple countries, data editing cannot be done by only one NSI. In those cases it is up to Eurostat, being responsible for European figures, to come up with the best possible solution.

2. Trust, but verify

a. **Statement**

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

b. **Rationale**

Successful data exchange between organisations is a shared responsibility of data producers and data consumers. This cannot be done without a reasonable amount of trust and understanding of each other's duties and challenges. It is the duty of data producers to validate data in the scope of the local perspective before providing it to others. It is the task of the data consumer to validate data in the scope of its broader perspective and provide data producers with useful feedback.

c. **Implications**

For the ESS this means that member States have a duty to provide Eurostat with data which conform to the validation rules agreed upon. Eurostat, guaranteeing and monitoring the quality of European statistics, has a duty to check that Member States data abide by these same rules and provide them with timely feedback on conformance.

3. Well-documented and appropriately communicated validation rules

a. **Statement**

Validation rules must be clearly and unambiguously defined and documented in order to achieve a common understanding and implementation among the different actors involved.

b. **Rationale**

This principle seeks (1) to facilitate the development of sound and efficient validation processes, (2) to formalise them and achieve their harmonised implementation and (3) to raise awareness of each participant's role in the validation process.

c. **Implications**

For the ESS two elements are needed to make this principle operational: a common and easy understandable validation language and an effective communication mechanism. This means that a universal validation language must be chosen by the ESS and that domain specialists (statistical working groups) must agree upon the validation rules for their respective domains.

4. Well-documented and appropriately communicated validation errors

a. **Statement**

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.

b. **Rationale**

This will ensure (1) that errors can be properly corrected, (2) their recurrence is minimised and (3) the risk of false negatives is reduced.

c. **Implications**

For the ESS this principle requires the definition of a standard ESS validation report structure that is expressive enough to explain the error, its type and severity at a minimum and clear and unambiguous enough to be easily understood by domain and data managers of the NSI's. A streamlined communication process between Eurostat and the NSI's is necessary to make this principle operational.

5. Comply or Explain

a. **Statement**

Validation rules must be satisfied or reasonably well explained.

b. **Rationale**

There may be situations that even earlier agreed validation rules cannot be satisfied. In that case there should be a possibility to escape from them, but only with a well described and understandable explanation that is accepted by the data consumer.

c. **Implications**

For the ESS this means the validation architecture should provide for a mechanism to explain the exceptional case of non-conformance and to define criteria to decide when an explanation is sufficient. Too strict criteria might become unworkable, too relaxed will not gain the quality improvements necessary. We advise to put together a set of best practices for explanations based on the use of this principle in practice. Repeated occurrences of non-conformance require joint re-evaluation of earlier agreed validation rules.

6. Good enough is the new perfect

a. **Statement**

Validation rules should be fit-for-purpose: they should balance data consistency and accuracy requirements with timeliness and feasibility constraints.

b. **Rationale**

It is well known and accepted that perfect data is a myth: errors always exist. The responsibility of the statistician is to manage them so that the final outcome represents a good compromise between all dimensions of data quality.

c. **Implications**

For the ESS this means that for the design of domain-specific validation rules in the statistical working groups one should look for the right balance between:

- Number of errors to be detected: detecting too many errors risks slowing down the process and makes it inefficient; detecting too few creates the risk that important errors are left undetected.
- Level of severity: rules that are too strict could slow down the process or may lead to a high rate of false positives.
- Level of complexity: rules that are too complex could be source of inconsistencies and therefore of flagging false errors.
- Output orientation versus book-keeping: validation rules should have a clear purpose in the broader context of the statistical output to be created.