UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

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

The ESSnet ValiDat Integration

Prepared by Volker Weichert, Federal Statistical Office of Germany

I. Introduction

1. The ESSnet ValiDat Integration [ESSnet (2017)] is a team effort of the national statistical institutes (NSIs) of the Netherlands, Portugal, Sweden, Poland, Lithuania and Germany to investigate the technical and methodological implications of establishing a common validation infrastructure for the European Statistical System (ESS). To reach that goal, three main topics of research have been identified: metrics, data formats and technical infrastructure.

2. The ESSnet builds on the results of the previous ESSnet ValiDat Foundation that was launched in 2014 to address the ESS' vision of harmonizing the validation approaches among its members. This ESSnet gathered and systemised information on the validation process of NSIs from all over the ESS. In the course of ValiDat Foundation, the common practice of data validation was described in a methodological handbook [Di Zio et al. (2016)] that was then extended to include some quality indicators and metrics for validation rules.

3. Furthermore, the ESSnet evaluated the Validation and Transformation Language (VTL) to determine its usability as a common validation language for the ESS [Gelsema (2015)]. While Eurostat has decided to use VTL for its own data validation, we have no knowledge of any NSI that followed that lead. However, there are plans to use the language for data validation in several NSIs, central banks and other organisations that frequently send data to Eurostat.[1]

4. Finally, ValiDat Foundation described a European infrastructure to store and distribute validation rules in VTL, and process the rules in a service infrastructure provided by the ESS. The ESSnet ValiDat Integration continues where ValiDat Foundation left off.

II. Data Validation Methodology

5. One of the main deliverables of ValiDat Foundation was a methodological handbook for validation. It is the result of a survey on the national practice of data validation in the ESS and of pioneer work on validation metrics.

6. Accordingly, the handbook consists of two main parts: one about data validation in general that explores the scope of validation rules, and one that focusses on metrics to help determine the quality of a rule or a set of rules. The first part also provides two ways of classification of validation rules: one formal and one practical.

[1] For simplicity we speak of NSIs in this paper. However, the same principles apply to other organisations that provide data to Eurostat, with the exception of matters that concern explicitly the ESS.

A. Improving the Handbook

7. One of the goals of the ESSnet ValiDat Integration is to improve the methodological handbook by including feedback gained from multiple sources. This feedback was originally thought to come from pilot projects to implement the handbook in some member states of the ESS. However, we identified several further sources of input to the handbook and will also incorporate feedback from an evaluation on VTL performed by Eurostat in 2016 and early 2017, validation manuals on national level, and domain-specific documents on validation.

8. In the end, the methodological handbook should be a user-friendly guide on how to perform validation. It should provide the reader with the means to determine the scope, complexity and necessity of a validation rule and understand what is entailed in its processing.

B. Metrics for Validation

9. While the theoretical background on data validation and the ability to gauge the complexity of a rule is necessary to design an efficient rule set and understand the implications of implementing it, it is also important to gain an understanding of its quality. This concept does not only refer to the validated data that is the result of the rule set applied to a set of raw data, but also to the rules themselves.

10. It is easy to understand that contradicting rules will cause the validation rule set to fail on every input. However, the avoidance of redundant rules, e.g., is not only a cosmetic issue, but also a question of performance. Other performance issues may be the result of rules that may deliver the intended result, but implement an inefficient way of getting there. To avoid these problems, a deeper understanding of the data structures and the validation language is necessary. Metrics that support the rule designer by identifying performance problems can be useful, either by making automatic removal of redundancies prior to execution possible or by pointing out the problematic rules.

11. Naturally, there are other aspects apart from performance that can immensely benefit from sound metrics that give a concrete indication of the quality of the rule set. Validation metrics derived from the output of the validation sub process can help identify systematic errors, problems in survey design, and errors not yet covered by rules. The rule set can be improved for the next iteration (see Figure 1). As part of the ESSnet ValiDat Integration, the portfolio of metrics in the methodological handbook will be reviewed, improved and extended.

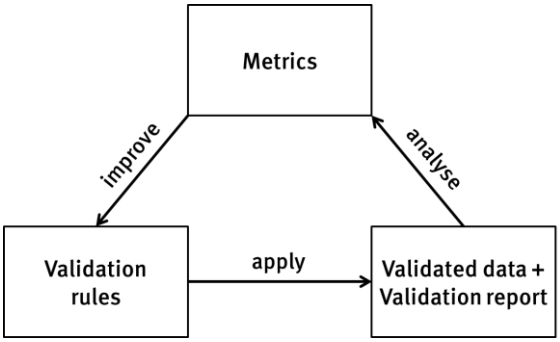

Figure 1: The role of validation metrics

III. Data Structures for Data Validation Reports

12. Data validation can be seen as a function $f(d_r, v) \rightarrow (d_v, r)$, where d_r is raw data, v is the validation rule set, d_v is the validated data and r is the validation report (see Figure 2).

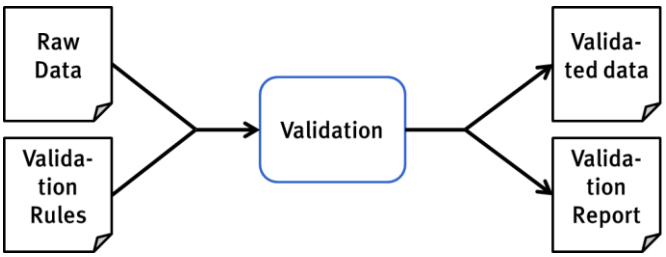

Figure 2: Input and output of the validation sub process

13. The validation report serves multiple functions. It is used as a communication device to tell the data owner that the data was either accepted or rejected, and indicate questionable or erroneous data in the set. When used to calculate metrics on the validation process, it is also a tool to tell the rules designer that the rule set may or may not be adequate.[2]

A. Machine- and Human-Readable

14. For versatility, the report should be available in a machine-readable version as well as in a human-readable version. This goal is often reached through applying visualisations like style sheets to the machine-readable format.

15. That way, the machine-readable format can contain a lot of data not necessarily important to every interested party, while different views on the report can be limited to data only targeted at specific audience.

B. Content

16. Using validation reports from different NSIs and Eurostat as a basis, we compiled an extensive list of information that could be useful in a validation report. The list contains data from categories like administrative data, operational data, metrics, warnings/errors and their specifics. In our next steps we will consolidate this information and then decide on a format to use for the report.

IV. Validation Infrastructure

17. A large part of the ESSnet ValiDat Integration is the continuation of the previous work on a common validation infrastructure for the ESS. The positioning of the validation sub process in a European statistical service architecture and its integration in the statistical production tool chain of an NSI is dependent on a lot of factors, many of which we probably do not yet know of.

18. The ESSnet's work on this consists of weighing the cost and the benefit of different strategies to enable an NSI's production system to implement validation rules coming from the ESS, either because they were shared by an NSI or because Eurostat requires data sent to them to be thus validated. It only makes sense to determine the feasibility, cost and benefits of different approaches by pilot implementations.

A. Background: Eurostat and VTL

19. Eurostat is provided data by NSIs, central banks, and other national authorities (ONAs) like e.g. ministries. While each organisation validates the data using their own rules, Eurostat also has its own set of validation rules to be applied to each data delivery.

[2] The metrics may already be calculated by the validation system and included in the report. The result is the same in either case.

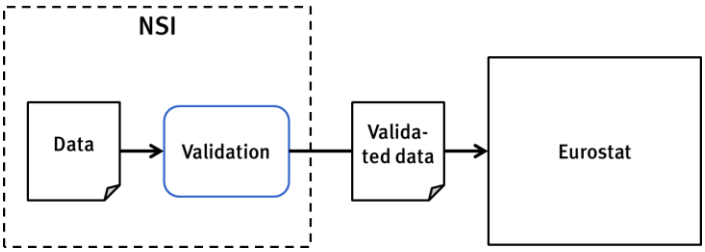

Figure 3: Validation is the NSI's responsibility

20. To improve the process of data validation for this setup, a common validation language (VTL) was developed. To facilitate a fluent exchange of data, a system has to be developed to enable each organisation to validate its data according to Eurostat's rules before passing them on (see Figure 3). This system should be versatile, so that each organisation can find a way to integrate the process into its own.

21. Just looking at the validation systems of the NSIs of the ESS, a wide variety of technical implementations can be found. While some NSIs have developed standardised tool chains for validation, using the same tools and languages for different domains and data sources, others have special tools for each domain, using different languages to express the validation rules.

22. The common validation system needs to present all NSIs with a solution that enables them to perform their validations for Eurostat while not impeding their own processes. The goal is to establish a service platform in the ESS, where validation services are provided either by Eurostat or NSIs. These services can be used by NSIs to perform the validation, but need not if it is not necessary.

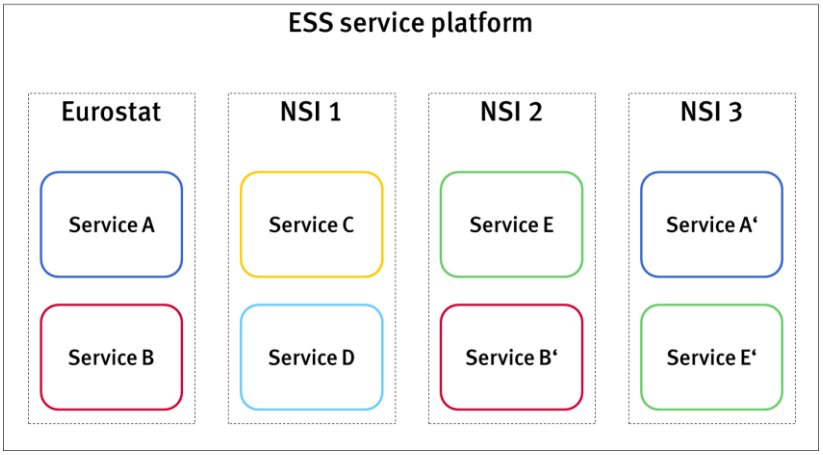

Figure 4: An example of a common ESS service platform

23. Figure 4 shows an example of such a service platform. Each member of the ESS can share their own services with the ESS, whose members can choose to use them in any way they prefer, shared or replicated. It is also feasible to replicate a service from another member and then share it again to improve overall system stability.

24 Two central services would be the rules and data structure registries that distribute the VTL rules and corresponding data structures to the ESS. Eurostat will provide a service bus to orchestrate the services, but each NSI is free to use their own implementation as well.

B. Validation Service Architecture

25. Three scenarios of integrating common validation services into national production systems are currently being considered by the ESSnet ValiDat Integration [Gramaglia (2015)]. The first scenario (see Figure 5) appears to be a natural choice for NSIs with a highly developed and standardised production system: while the rules and data structures are provided by servers in the ESS service platform, the entire validation process is performed on the NSIs own system.

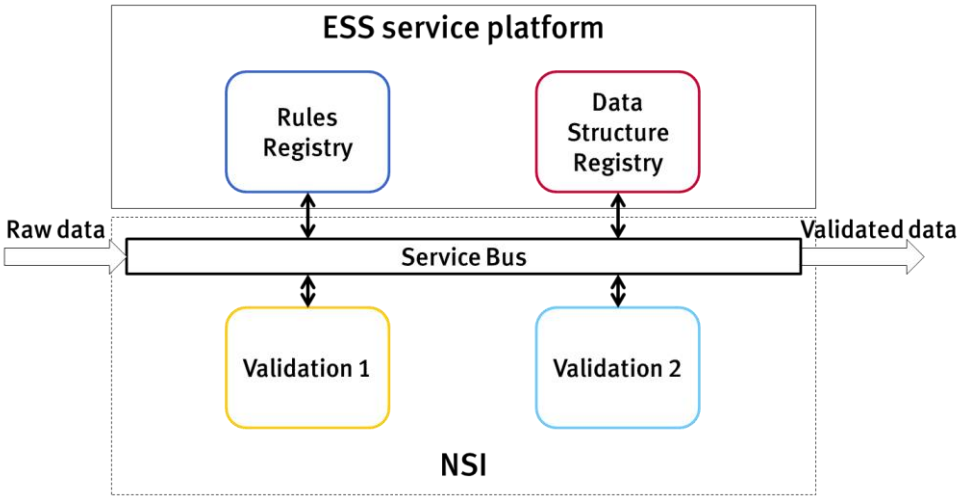

Figure 5: Scenario 1, NSI uses only rules and data structures from ESS platform

26. The second scenario makes more use of the service platform. Again, the rules and data structures are retrieved from the ESS service platform, as well as some parts of the validation process that are performed using services provided by the ESS (see Figure 6). That way, the process orchestrator would still be the national production system. However, some logic would not need to be included in the national system if it were already provided by the ESS.

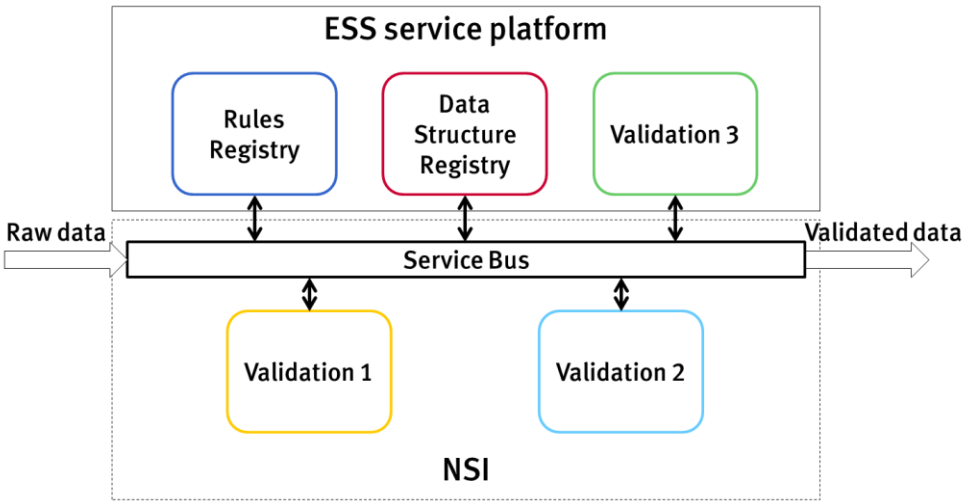

Figure 6: Scenario 2, NSI uses registries and some validation services

27. In the final scenario, the validation sub process is completely handled by the ESS service platform. The national system hands the data to Eurostat's process orchestrator on the service platform and receives the validated data and validation report at the end (see Figure 7).

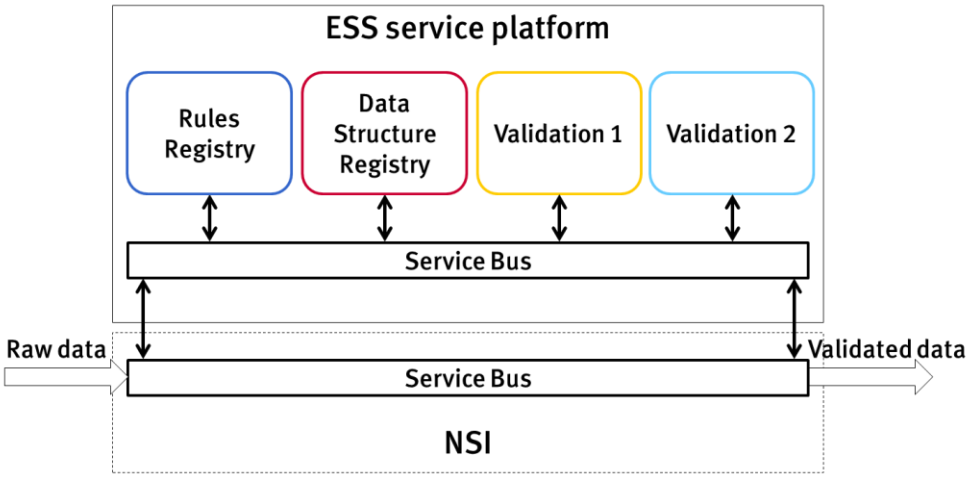

Figure 7: Scenario 3, the entire validation sub process is handled on the ESS platform

D. VTL converter

28. The first two scenarios rely on the national system to perform at least some of the validation. However, the rules on the rule registry are provided in VTL. Therefore, a cross-compiler from VTL to the national validation language needs to be developed in order to implement these scenarios.

29. We will include a study on the feasibility of such a cross-compiler based on a pilot implantation of a converter from VTL to SQL. This implementation will be the result of the continued development of the converter of the Regional Statistical Office in Olsztyn [Olsztyn (2017)].

E. Cost-Benefit-Analysis

30. One of the main goals of this project is to determine a way to advise NSIs on the way to integrate the common validation services into their own production systems. In order to do that, we need to be aware of the consequences that each choice brings.

31. We are preparing an analysis of the cost of each integration scenario, and its benefits. The validation infrastructures of NSIs differ considerably in organisation and implementation and this diversity affects the perceived and actual costs and benefits of different solutions.

32. In a first step we will try to cluster the different validation architectures into groups of similar solutions. This classification will depend on many factors like the degree of centralization and automation, the progressiveness of the IT-infrastructure, the expertise of the available personnel and the conformity to European standards.

33. E.g. if an NSI has already built an elaborate and standardised national validation system, a solution at interface level promises minimum impact on the productive system, and would seem a better fit than a solution based on using the new ESS-based validation services. However, these are but two of the aspects taken into consideration by a thorough cost-benefit analysis (standardisation and technical progressiveness).

34. Another important point to consider is that the benefits may not only apply to the delivery of data to Eurostat. The validation system provided by the ESS platform may also be used for the internal validation processes of the NSIs, if the rules (in VTL) and the data structures are available on the corresponding servers.

35. We will use two main sources of input to develop the classification of different validation system types. Eurostat is currently working on a framework (MSBM) for categorisation of NSIs based on the UNECE Modernisation Maturity Model [Dalton et al. (2015)] and a more empirical model has been developed from data collected in a validation survey in the ESS in early 2017. Data obtained from the survey and the categorisation framework point in the same direction: the integration and standardisation

of validation systems depends on the combination of methodology, organisation, and IT-infrastructure and ignoring even one of these aspects will result in unsatisfactory results.

36. The second step will be pilot implementations of the three scenarios using existing services like Eurostat's *struval* and *conval*, new services that will be developed by the ESSnet, and other modules necessary to fulfil either the new European policy, or advancing the national production. These concrete, but typified solutions will represent the envisioned final state of validation systems.

37. Next, we will model the steps necessary for archetypical NSIs to develop their validation systems to the point where they resemble one of the three solutions. Due to economic factors, some of the paths from old to new system will not seem feasible and therefore not be explored in this part of the ESSnet. Drawing on the experience from the pilot implementations, the main contributors to the cost of each remaining path will (qualitative and quantitative) be identified and estimated. Of course, for some factors we will have to rely on educated guesses.

38. We will build a multi-criteria model including those cost drivers. This model will also take into account that soft factors might have a great impact on an NSI's decision to follow a given migration path. We expect a rather large number of estimated and thus subjective factors to be included in the model. Therefore, a sensitivity analysis needs to be performed to make sure that individual estimation errors do not affect the outcome decisively.

39. The resulting model should provide NSIs with an orientation point for the decision on the most suitable scenario: first they determine the type of their existing national validation system. Then they apply individual weights to the different cost factors and benefits each solution provides, and finally they follow the path that leads to the most desirable solution in terms of cost and benefits. A simple example is given in Figure 8.

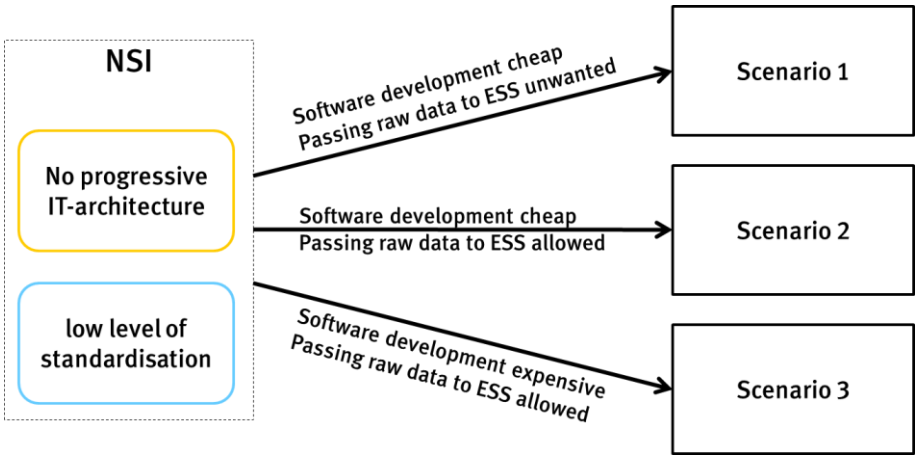

Figure 8: Multi-criteria analysis using two characteristics and two cost drivers

V. Conclusion

40. International cooperation on validation is a difficult topic. Many different national systems in very diverse states of development hinder the process of standardization and thus the exchange of knowledge (rules) and experience (metrics). The ESSnet ValiDat Integration is aiming to level the field a bit by exploring ways to not only develop a common validation platform in the ESS, but also make it integrable into national statistical production systems.

References

ESSnet ValiDat Integration (2017), CROS portal, https://ec.europa.eu/eurostat/cros/content/essnet-validat-integration_en.

Di Zio, M., Fursova, N. Gelsema, T. Gießing, S. Guarnera, U, Petrauskienė, J., Quensel-von Kalben, L., Scanu, M, ten Bosch, K.O., van der Loo, M. Walsdorfer, K. (2016), *Methodology for Data Validation. Revised edition*. ESSnet ValiDat Foundation.

Gelsema, T. (2015), *A study of VTL*. ESSnet ValiDat Foundation.

Gramaglia, L. (2015), *Towards a European validation architecture*, ESSnet ValiDat Foundation Workshop, Wiesbaden.

Olsztyn, Regional Statistical Office in (2017), *Technical Aspects of VTL to SQL Translation*. UNECE Work Session on Statistical Data Editing, Den Haag.

Dalton, P., Dunne, J., Kelly, D. (2015), *Modernisation: Evolution or Revolution*, Global Conference on a Transformative Agenda for Official Statistics, New York.