Common Validation Policy – A Member State Perspective

Lucas Quensel-von Kalben[1]

[1] *Destatis,Wiesbaden, Germany;lucas.quensel-von-Kalben@destatis.de*

Abstract
A common validation policy within the European Statistical System is of paramount importance both from a quality and a productivity view. Since 2012 Eurostat is working on such a solution taking methodological, technological and infrastructural issues into account. The National Statistical Institutes contribute in various ways to this work. This involvement should be extended to secure that these solutions can be handled by NSI staff, is adapted to national production environments and fulfill the expectations in terms of future quality and efficiency gains. The paper addresses some major issues and motivates national institutes of the member states to a more active participation.

Keywords: ESS Vision 2020, Validation, VTL, Shared services

1. Introduction

Validation matters. This is the first fundamental of the Wiesbaden Manifesto (Wiesbaden, 2015), a collection of results of an international workshop held at Destatis 2015 on Validation policy.

But what is “data validation”? According to the UNECE, it is "an activity aimed at verifying whether the value of a data item comes from the given (finite or infinite) set of acceptable values." (UNECE, 2013). Data editing and imputation are separate but closely related activities. In the real world these activities are not always kept apart which led occasionally to methodological confusion.

In 2012 Eurostat started an initiative on harmonizing data validation policy. It set up a program of interconnected projects that are now part of Vision 2020 Implementing bodies and projects (Steering Group, ESS.VIP, Task Force, ESSnet).[1] The whole structure of governance and projects will not be differentiated further in this paper and referred as the "validation project".

Two goals were propagated by the validation project:

a) Improve the transparency (a quality issue) of validation processes across institutional boundaries
b) Increase the interoperability (a productivity goal) by fostering common technical solutions

From Eurostat perspective the main focus of the project is on the interface between the NSIs and Eurostat, i.e. validation of data sent to Eurostat. This will be defined as the "narrow focus". The "wider focus" is a different view that includes the whole statistical production chain from data collection at national to dissemination on European level. The differences in perspective and its consequences will be discussed in this paper. They apply mainly on the organizational and technical side and are less relevant for methodological and language aspects.

2. Common Validation

The Vision Implementing Project and the ESSnet worked in a logical sequence of steps from requirements to solutions. Each step has its own deliverables. They will be used as starting point for discussion.

2.1. Validation in the European Statistical System

Eurostat started early to analyze validation practices within its own subject matter departments. Two main results could be identified: (1) Validation and its documentation are not well defined nor is the process of its specification in the ESS (2) Data transmitted to Eurostat from the NSIs (and other national authorities) do often not fulfill the expectations of

[1] The VIP and the ESSnet have finished their work at the end of 2015. At the time of writing a new project was on its way, starting in late 2016.

Eurostat in terms of quality. This results in transmission “ping-pong” until the data are finally accepted.

The situation in the ESS member states is even more heterogeneous as a survey conducted by the ESSnet in spring 2015 discovered. In this survey the NSIs in general and five subject matter domains (Census, Agriculture, Prices, Labour Force and Structural Business Surveys) in particular were asked about organizational, methodological, technical and productivity aspects of their validation processes.

The main findings of the European survey on validation have been presented several times (Gießing, 2015a; Gießing 2015b). To highlight some results: In almost all institutions no common validation rules and standards of specification exist, the same applies to validation rules across different offices within the same domain. Validation is a process which spreads across all four production phases of GSBPM (data collection, preparation, analyzing and dissemination). The organization of validation is not standardized and most offices follow a decentralized approach where validation is defined and used in the subject matter departments only. The usage of IT-applications, tools and services is not harmonized at all. General purpose software products like MS-Excel, SAS or direct manipulation within databases with SQL are most common. Specific validation tools are used in some national institutes.

The level of “maturity” of validation in most member states and Eurostat is still optimizable. This observation is surprising when compared to the amount of work spent for validation (and its - methodologically speaking – sister processes data editing and data imputation). The effort spent was estimated between 40 and 60 % of the overall workload.

Two conclusions could be inferred for the goals of the project from member state perspective. First, there is a real issue with validation in the ESS. Second, national and international starting points are very different and need to be included into new solutions. A “one size fits all” approach is clearly not applicable.

2.2. Methodological Foundations

Starting from current state of validation in the ESS, the project developed a methodological framework, "the handbook", to communicate concepts and a common vocabulary for the ESS (De Zio et al., 2015).[2] The handbook is not a master template for individual concepts in specific domains. Its main purpose is to define validation in terms of the big "W"-questions. What is validation? Why validate data? When and how to apply validation? How to improve validation in the validation lifecycle? How to "measure" the impact of validation in statistical production (how much effort should be spent)?

Some content is probably of more value to the NSIs then other. In the survey on validation in the ESS, we faced major difficulties with a clear typology (or classification or levels) of validation that could be used as lingua franca and would be easily understandable by the statisticians in the ESS and beyond. The handbook suggested different "schemes" **(Fig. 1)** and its interrelationship. From formal to rather pragmatic classifications, a clearer understanding of the validation process can be achieved.

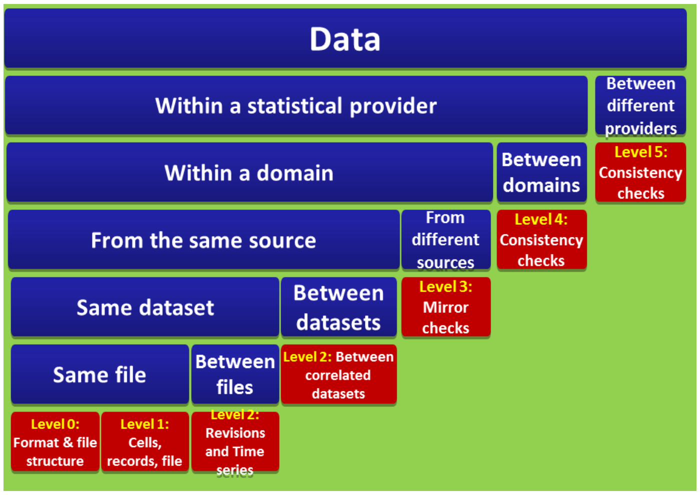

Data
Within a statistical provider
Between different providers
Within a domain
Between domains
Level 5: Consistency checks
From the same source
From different sources
Level 4: Consistency checks
Same dataset
Between datasets
Level 3: Mirror checks
Same file
Between files
Level 2: Between correlated datasets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

Fig. 1 Validation levels (Simon, 2013)

[2] A similar approach has been started at UNECE-level by a working group developing a "Generic Statistical Data Editing Model" (GSDEM). The scope of GSDEM is wider than the scope of the validation handbook and includes other methods of error detection and data editing and imputation. Both (GSDEM and handbook) complement each other well when the vocabulary is harmonized.

Another interesting aspect is metrics for sets of validation rules. Questions of optimized rule sets and how to detect them, have not been answered fully yet but some ideas have been launched.

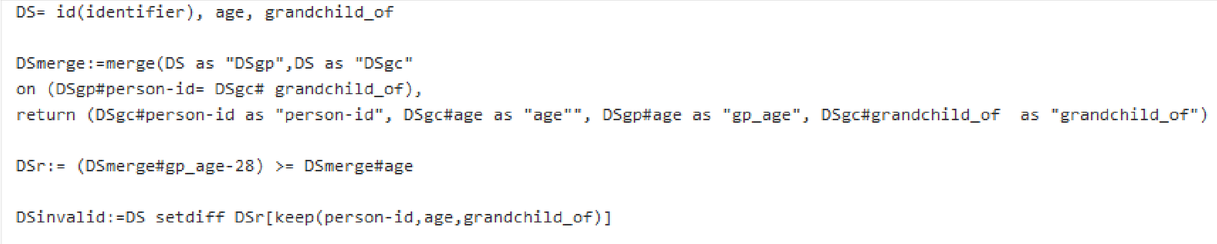
How to deal with the handbook on national and international level? A starting point would be a discussion of the concepts and a reflection of the processes being used now. This should be done as a joined endeavor of subject matter specialists, methodologists and IT-people. The discussion itself could be worthwhile to develop a common understanding and awareness of need for harmonization.

2.3. Language and standardization

Parallel to the work being done in the validation project, the SDMX-community, i.e. an international and independent body, was and is working on a new standard language for specifying validation rules and more. This so called Validation and Transformation Language (VTL) has been published in its version 1.0 in spring 2015. VTL is meant as an internationally agreed language which is at the same time human readable and formally enough to be used as input for machines to evaluate data against validation rules.

The validation project evaluated VTL from different perspectives. Questions like the usability and the machine readability, the coherence and functional completeness have been addressed. The critique of the project regarding maturity and adaptiveness to purpose was quite substantial (Geselma et al., 2015). Some improvement actions were proposed. The version 1.1 (public available in June 2016) should adapt to these improvement actions.

A further experiment was launched within the validation project (internally termed proof-of-concept – PoC) (Van der Loo, 2015). 18 Rules were chosen and specified in natural language. Nine rules were taken from the ESS survey and nine specified as proxies for general classes of validation rules from a more abstract point of view (covering the theoretically possible dimensions of validation rules in general and thought to be “complete”). These rules were translated into VTL (**Fig. 2**). In a further step the VTL’ized rules were taken as input for two “national” validation languages. The Dutch specification language is based on an application in R. The German specification is used in the German Statistical System. Finally test data have been checked against the rule sets. Several issues have been raised by the experiment. The good news is that all rules could be specified in all three languages, albeit with some tricky work around both in VTL and PL-Spezifikationssprache, the German language. The languages are “complete”. Other criteria were less successful met. A transformation from VTL to national languages faced heavy difficulties. The resulting code from VTL was neither intuitive understandable (the human perspective) nor automatically translatable (the machine readability).[3]



```
DS= id(identifier), age, grandchild_of

DSmerge:=merge(DS as "DSgp",DS as "DSgc"
on (DSgp#person-id= DSgc# grandchild_of),
return (DSgc#person-id as "person-id", DSgc#age as "age"", DSgp#age as "gp_age", DSgc#grandchild_of  as "grandchild_of")

DSr:= (DSmerge#gp_age-28) >= DSmerge#age

DSinvalid:=DS setdiff DSr[keep(person-id,age,grandchild_of)]
```

Fig. 2 Example of validation rule in VTL (Van der Loo, 2016)

2.4. Organization and Process of validation

The project suggests a new kind of collaboration between Eurostat and ESS member states. The collaboration (**Fig. 3**) starts at the design phase when validation rules will be discussed and agreed upon together. Communication should use the vocabulary and concepts of the methodological handbook to facilitate a common understanding between experts. A more formal specification will then be provided in VTL. The specified rule sets will be uploaded to a central registry and being used by specific IT-services (either centrally provided or in the national production environments) to validate data of a specific domain.

[3] The test identified some problems with the German specification language as well. The Dutch code was the most concise and elegant. It was also probably the best to write and understand.

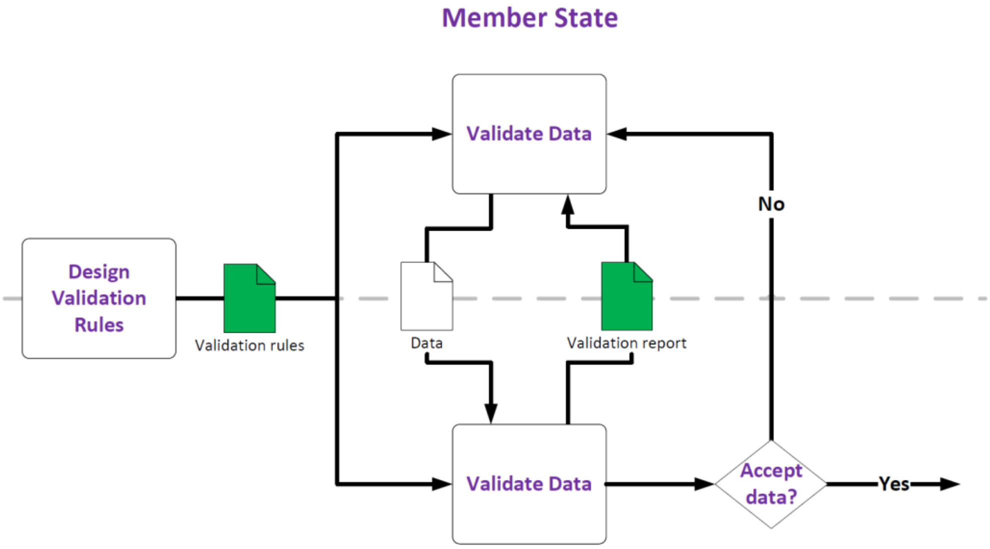


Fig. 3 Process flown at transmission to Eurostat (Gramaglia, 2016)

The distinction between a narrow and a wider focus has already been introduced. The scenario just developed refers to the narrow focus. Taking into account that the major workload in validation takes place in national production, the benefits could be increased significantly by incorporating national validation rules to the registry and using the services for national purposes as well.

2.4. IT-Architecture: Tools and Services

Eurostat developed some first prototypes of tools and services which give a glimpse of a future IT-architecture in Europe for validation. The “validation system” comprises three main components (**Fig. 4**). A repository for the validation rule sets is the center of the system. A rule editor/builder should ease the handling of VTL and hide some of its complexity. It is planned as a Graphical User Interface (GUI) application that can be used to create and edit validation rules interactively. The third component is IT-services that use validation rules created by the GUI-application and stored in the central repository. Currently two sub-components are/have been developed by Eurostat as prototypical solutions. One is used for “structural” validation, i.e. validation rules that check compatibility to more structurally oriented rule types (data structure, file format and basic checks on attribute lists). This service, StruVal, is based on the SDMX-converter and limited to data prepared in this format.

The second IT-service has just been started in development and focuses at the validation of content. The service should be able to interpret VTL and check data against it. The logic in ConVal (content validation) is far more complex than the structural validation service. It will be interesting to see the results.

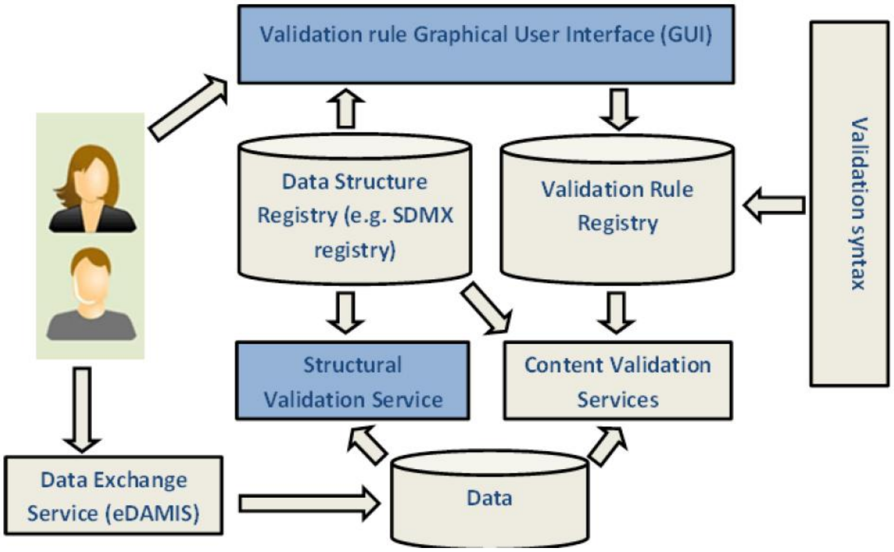

Validation rule Graphical User Interface (GUI)
Data Structure Registry (e.g. SDMX registry)
Validation Rule Registry
Validation syntax
Structural Validation Service
Content Validation Services
Data Exchange Service (eDAMIS)
Data

Fig. 4 IT-Architecture of validation tools and services (Gramaglia, 2015)

Services and registry are centrally hosted at Eurostat. While this is probably the best solution for the registry, it might not be the best for the validation services. Most member states would prefer services which can be implemented within their own premises. This is not just a political issue but can be explained with objective reasons as well. Questions of data protection and IT-security, performance and stability can be easier solved by sending rules from a central registry to local services than sending data to centrally hosted services. Using such services remote (at Eurostat) is cheaper in terms of implementation and maintenance costs on the other hand. A secure and performing network is a prerequisite for a central solution. The choice of a local hosting scenario is even more important when the services are used for national validation purposes as well. Even replicas of the central registry or extended versions might be necessary because of the vital significance of availability of services and rules sets for national production.

The wider focus of integrating services and tool in national production is no easy endeavor. Solutions will differ according to the national production infrastructure already existing. NSIs that are using many different all-purpose tools can replace these tools bit by bit with the new services.[4] A different approach is more applicable to member states that have invested much into heavy weight production chains in the past.[5] Here adapters between European and national systems could be more applicable.

Narrow and wider focuses are not necessarily in conflict with one another. A stepwise approach from the narrow to the wider focus should reduce risks and increase familiarity with the concepts.

3. Conclusion and recommendations

Validation seems on first glance a rather dull subject compared to some other topics of the ESS. It is very basic to statistical production , consumes a lot of energy and is paramount to quality improvements. The maturity of this process is still in its infancy which is in strong contrast to its significance. Dealing with an improvement of this process, promises quality and efficiency gains in medium terms.

Most aspects have been discussed in detail in the validation workshop in Wiesbaden November 2015. The manifesto (Wiesbaden, 2015) is a shortcut to the main conclusions.

From the conclusions several recommendations for the member states can be stated.

1. Take validation and the developments on European level serious. Appoint members of staff as contact point for further involvement and feedback. Think about participation in Steering groups, Task Forces, ESSnets and other bodies
2. Make your offices familiar with the concepts and vocabulary of the European validation world. Adopt these in your own institutions. Attend training sessions (ESTP) and network

[4] Another ESS project (ESS.VIP SERV) is actually trying to standardize the way services can be used. The project is based on an emerging standard for the interoperability of services in official statistics (Common Statistical Production Architecture – CSPA). This standard is promoted by the High level group on the modernization of Official Statistics (HLG MOS).

[5] The German Statistical System already has a similar infrastructure as the one proposed by Eurostat for the ESS. Here the rules sets (as other metadata necessary for statistical production) are stored in a survey database. This registry supplies services and individual applications at different phases of production with the required rule sets.

with member states actively involved in the validation project (in lieu of a Centre of Excellence)

3. Analyze VTL as the most likely language for specifying validation rules
4. Explore tools and services provided by Eurostat. Give a feedback for improvement. Check for implementing tools and service or some kind of intermediate layer in your own production chain.

The participation in these activities is comparably cheap and has a high impact on quality and productivity.

4. References

De Zio, M. et al. (2015), A generic framework for data validation (Work Package2 deliverable, ESSNet on Validation)

Gelsema, T. et al. (2015), A study of VTL (Work Package 4 deliverable, ESSNet on Validation)

Gießing, S. and Walsdorfer, K. (2015), Validation in European official statistics: Results of an ESS survey, https://ec.europa.eu/eurostat/cros/sites/crosportal/files//2%20-%20ESSnet_ValiDat_survey.pptx

Gramaglia, L. (2015), Towards a European validation architecture, https://ec.europa.eu/eurostat/cros/sites/crosportal/files//1%20-%20Wiesbaden%20Workshop.pptx

Gramaglia, L. (2016), ESS Vision 2020 Validation project: deployment actions, Item 2c of the agenda of the 29th Meeting of the European Statistical System Committee

Simon A., (2013), Definition of validation levels and other related concepts v01307. Working document; https://webgate.ec.europa.eu/fpfis/mwikis/essvalidserv/images/3/30/Eurostat_-_definition_validation_levels_and_other_related_concepts_v01307.doc

Van der Loo, M. et al. (2016), A VTL proof of concept (Work Package 4 deliverable, ESSNet on Validation)

UNECE (2013), Glossary of terms on statistical data editing, http://www1.unece.org/stat/platform/display/kbase/Glossary

Walsdorfer, K. et al. (2015), Consolidated response of the survey on validation rules (Work Package 1 deliverable, ESSNet on Validation)

Wiesbaden (2015), Manifesto from the Wiesbaden Workshop on Validation, https://ec.europa.eu/eurostat/cros/sites/crosportal/files//Manifesto%20from%20Wiesbaden.pdf