The ESS.VIP Validation and its implementation in waste statistics

Hartmut Schrör and Ángel Simón Delgado, Eurostat

The aim of the ESS.VIP on Common Data Validation Policy is to deploy a coherent validation policy in the different statistical domains in cooperation with Member States. This policy will include in particular a distribution of validation tasks along the production chain starting in Member States and ending at Eurostat.

Waste Statistics, as reported to Eurostat according to the Waste Statistics Regulation (WStatR), are a complex matter for statistical measurement. Following the recommendations of the ESS.VIP, Eurostat has therefore started formalising data validation with our data providers in the Member States. The expected benefit is to have a validation system that is more efficient both for Member States and Eurostat in Waste Statistics and thus to improve the quality and timeliness of the disseminated data.

**1. The ESS.VIP Common Data Validation Policy**

*1.1 Why an ESS.VIP on Validation?*

In 2012, Eurostat launched the ESS.VIP Programme, which is a set of 'Vision Implementing Projects' with the objective to put into practice key elements of Commission Communica­tion COM(2009) 404 on the 'Vision for the next decade' in the ESS.

One of the projects is the ESS.VIP Common Data Validation Policy, referred to also as the ESS.VIP Validation[[1]](#footnote-1). Its objective is to deploy a coherent validation policy in the different statistical domains in cooperation with Member States. Thus, the project is a response to the problem that validation is often carried out in Member States and at Eurostat in an uncoordinated way. The effect may be duplication of work on the one hand or validation gaps on the other hand. In addition, validation is not described and documented in a standardised way, thus making it more difficult than necessary to know about duplication and gaps and to communicate validation rules effectively. Finally, a variety of tools is used for validation in the ESS, all of which are designed to serve similar purposes. As a result, validation in the ESS is not as efficient as it could be and would benefit considerably from a more coordinated and collaborative approach. The expected benefit would be a more timely dissemination of higher quality data at lower costs both for Member States and Eurostat.

*1.2 Principles of the ESS.VIP Validation*

The project pays special attention to both horizontal integration (between domains) and vertical integration (between Member States and Eurostat):

* Horizontal integration between business and horizontal units is to be achieved through the development of a common language for validation rules (common terminology for the description of validation levels, a common validation syntax) and its use, together with common validation tools in the statistical domains.
* Vertical integration between Member States and Eurostat is to be achieved through sharing validation rules in the common validation language and distributing valida­tion responsibilities, ensuring the coherence between data files and the integrity of the data held in the ESS.

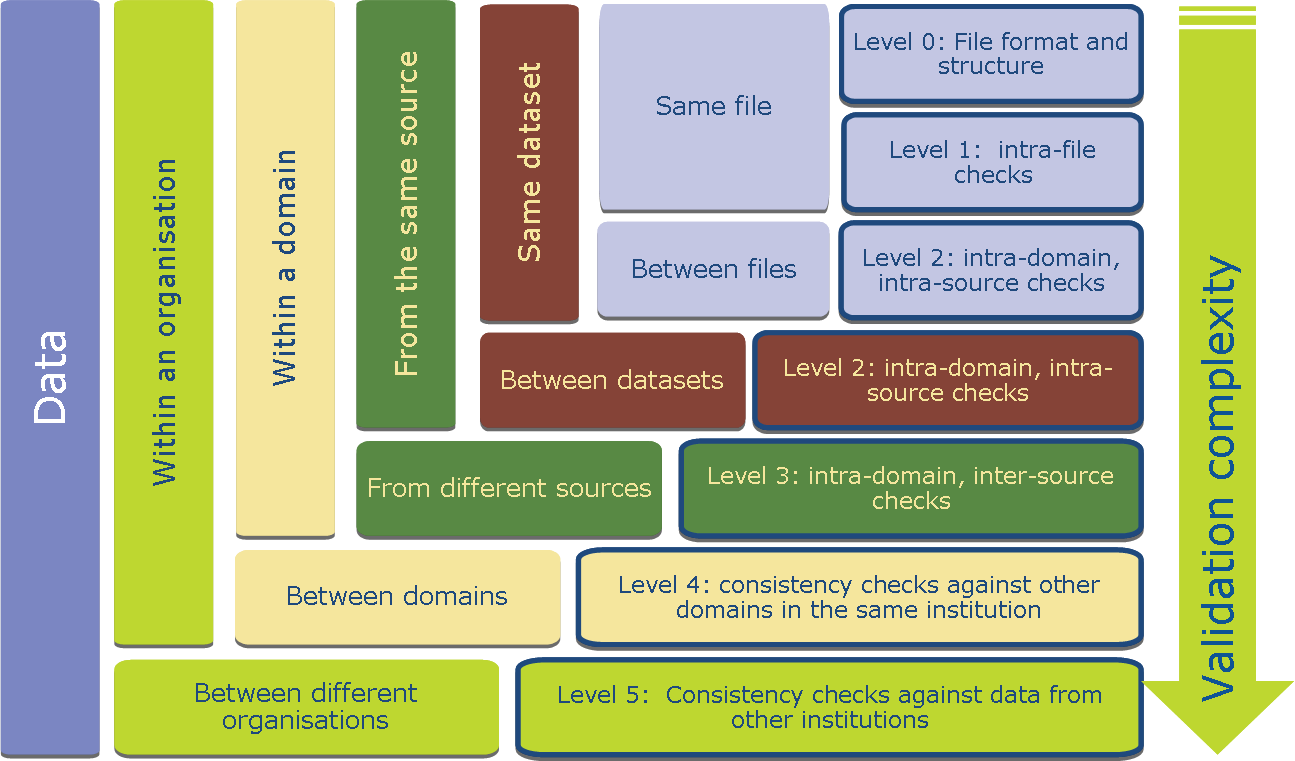
Concerning vertical integration, Eurostat's idea is to agree with the Member States by statistical domain on how the validation procedures, once defined and described, should be shared in the production chain at national and European level. The basic principle is "validation, the sooner the better". Eurostat's reasoning underlying this principle is not to allocate as much workload as possible to the Member States, but to make validation as a whole as efficient as possible. When Eurostat validates national data and asks Member States to revise or explain outliers, several months may have passed since the data compilation in a Member State. After this period of time, analysing and correcting or explaining data is often more difficult than it would have been earlier during the production process in the Member State. In addition, the lack of a common language describing validation rules and the resulting observations sometimes makes it difficult in Member States to respond to Eurostat's requests. Earlier validation in the production chain, based on a common understanding of the validation rules, would be an efficiency gain both at national and European level because less communication and revision at a late stage in the overall production process would be necessary.

*1.3 Activities of the ESS.VIP Validation*

Eurostat launched the ESS.VIP Validation in the fourth quarter of 2012, following an internal "VIP Validation", in which first conceptual work had been done concerning the standardisa­tion of validation rules and the design of a generic software tool for validation.

One result of the VIP Validation is the concept of the six validation levels, ranging from the most simple format and file structure checks to complex consistency checks across statistical domains and even across different organisations.

Figure 1: Validation levels and typology



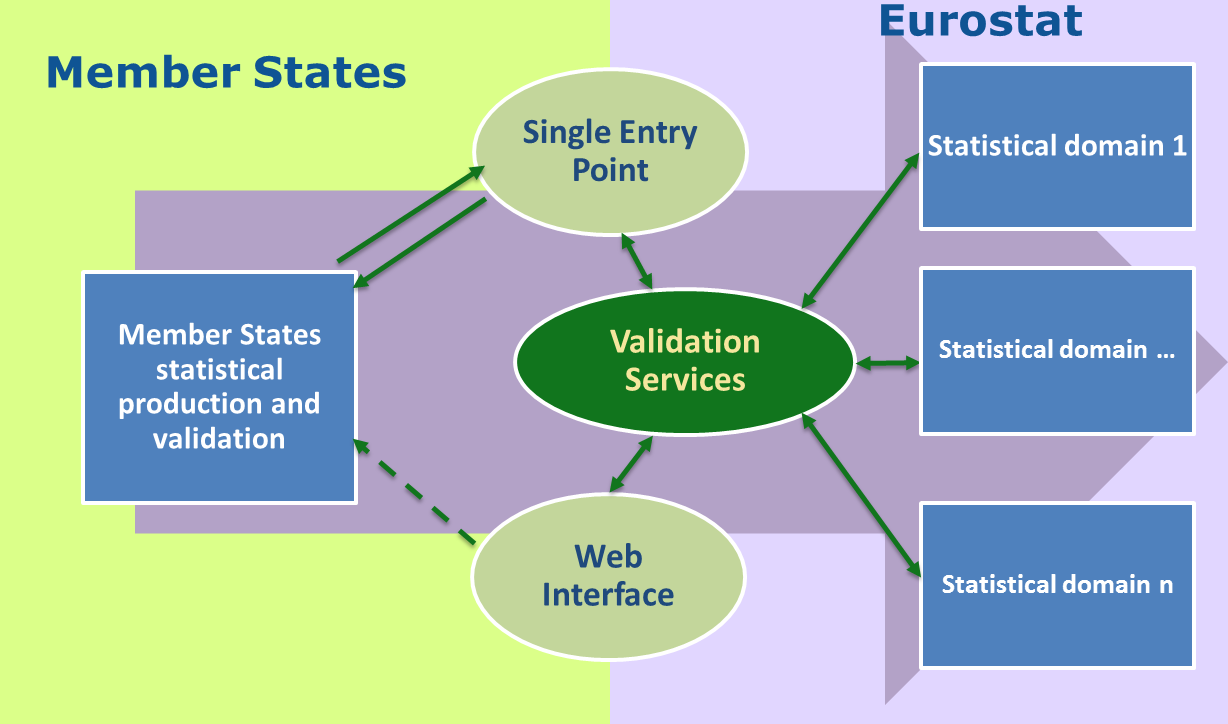
The development of a generic Validation Syntax (VALS) is on-going, which will permit the expression of all validation rules in a standard machine readable format. This will allow the storage and maintenance of validation rules in a common repository, which can be accessed by production systems performing validation checks. VALS has been chosen as one of the inputs, together with the EXL language from Banca d'Italia, for the development of the Validations and Transformations Language (VTL) in SDMX.

In addition, functional requirements for the tools and services of a validation architecture are being drafted, including a graphical user interface which will guide statisticians writing and editing validation rules in an intuitive way.

During 2013, Eurostat performed an internal assessment of the maturity level of validation in all statistical domains. The purpose was to create a catalogue of best practices as well as to identify the most common issues that hamper efficient validation and to develop proposals for a desirable maturity level.

As a longer term goal, Eurostat intends to implement a shared validation system with the Member States, which will offer a general validation service using web technology and a user-friendly interface for statisticians.

Figure 2: Elements of a validation system



The integration with the Single Entry Point at Eurostat, currently EDAMIS, will make it possible to call validation services in the process of transmitting data from a Member State to Eurostat. The validation rules would be stored in a common repository able to communicate with the environment using the common validation syntax VALS.

As the target validation architecture would extend to the ESS as a whole, a task force with five Member States was launched in February 2014, and the ESS.VIP Validation will obtain a broader basis with an ESSnet that is starting its work in the fourth quarter of 2014. The purpose is to benefit from experience with validation systems in Member States and to ensure that a common validation system responds to the needs of the Member States in the future.

**2 Validation in practice – the case of Waste Statistics**

*2.1 Waste Statistics: a developing statistical domain*

The Waste Statistics Regulation (WStatR)[[2]](#footnote-2) governing the harmonised data collection on waste generation and management entered into force in 2002. As the reporting frequency is biennial, data only on the four even reference years from 2004 to 2010 are currently available.

During the first decade of data collection, very detailed and profound validation of the national data has been essential to the quality, and comparability in particular, of waste statistics for several reasons.

First, the use of NACE Rev. 2 from 2008 onwards required recoding and verifying the previous data reported according to NACE Rev. 1.1. Then, a revision of the WStatR introduced a first break in several time series at detailed level from 2010 onwards. Moreover, methodology and experience with the data collection have been built up gradually over the first four data collections. Breaks in time series are sometimes due to changes in methodology.

Last but not least, waste generation and management are very complex subjects for statistical measurement. Data on waste generation are broken down into the sources of waste according to the NACE classification. As the sources cover not only the business economy (NACE B to N) but also agriculture, the public sector and households, usually various methods are combined to collect the data. Waste management is difficult to measure as well because it is often a multi-stage process with material flows that are sometimes unknown in detail. Looking at the breakdown into 51 waste materials, there is often a mismatch between the amount of waste generation and the corresponding treatment because of weight losses, classification errors, import and export of waste, storage effects, use of different data sources and double counts (both intentional or unintentional). Finally, although a detailed recommendations manual on waste statistics[[3]](#footnote-3) has been developed and first released in 2010, Member States can use various data collection methods because the detailed methodology is a matter of subsidiarity.

*2.2 Purpose of validation*

The assumption is frequently made that the purpose of validation is to detect and correct errors. This process is then automated as much as possible using software tools. While this applies to waste statistics as well, the second, equally important purpose has been to find out about outliers and peculiarities in the national data that reflect a diverse and volatile reality or methodological differences rather than a statistical error. For instance, the abrupt rise of the amount of waste incinerated in a small country may be due to one new waste incinerator starting its operation. Thus, the purpose of validation particularly in waste statistics is not only to have accurate, comparable, coherent and timely results, but also to have high quality reference metadata explaining peculiarities in the datasets to the users of our statistics.

*2.3 Vertical inegration - making validation of waste statistics more efficient*

Given the complexity of validation in waste statistics and the considerable workload involved both at Eurostat and in the Member States, the waste statistics team at Eurostat decided early to implement vertical integration to make validation as efficient as possible.

The first step in 2013 was to *take stock of the existing validation procedures* in Member States*.* For this purpose, we conducted a short survey among our data providers. They were asked to indicate on a checklist which validation rules they performed on their datasets before submission to Eurostat and, where appropriate, to add validation rules to the list that had not been considered in the survey.

As a second step, we tried to *identify a standard set of validation rules* that all Member States could equally apply to their data before submission to Eurostat. The purpose was to implement the principle "validation, the sooner the better" to ensure the desired level of maturity before data enter Eurostat and thus to reduce the requests for clarification or revision that Eurostat would send to countries afterwards.

To *discuss validation in detail*, we dedicated a workshop in September 2013 to the subject of "Validation in Waste Statistics". The questions that we addressed were "What validation do Member States perform and how?" "What validation rules are most useful, and how do we know?" "Is Eurostat's validation useful for data providers in Member States?" and finally "Can we identify a common set of validation rules to be applied in all Member States?"

As the results of the workshop were promising, Eurostat elaborated a *detailed description of standard validation rules* as a reference, which was welcomed at the next Working Group meeting on Waste Statistics in April 2014. Eurostat invited Member States to apply the agreed validation to the next data collection that is due by the end of June 2014 and to report on their observations in section 12.4 "validation" of the quality reports according to the ESQRS[[4]](#footnote-4). As the development on the validation syntax VALS was not finished by that time, we chose an informal notation for the rules. The example below shows the notation of one of the rules saying that amounts of treated waste should not exceed the respective treatment capacities reported by a county.

##### 4. Treated amounts vs. treatment capacities

The quality and comparability of data on waste treatment and of the waste treatment infrastructure shall be assessed by the comparison of the quantity of waste treated with reported capacities of treatment facilities for the operations energy recovery and incineration.

Short description:

* **Assumption**: Total amount treated equal or lower than available capacity
* **Data level:** Test applied to total amount (hazardous & non-hazardous waste) treated by:
  + Incineration with energy recovery (R1)
  + Incineration without energy recovery (D10)
* **Calculations:**
  + Calculation of the ratio of the total amount treated by energy recovery and incineration divided by the capacities of the respective treatment operations:  
      
    
* **Thresholds:**
  + lower threshold: None
  + upper threshold: ratio treatment/capacity > 1.0

*2.4 Criteria for selecting standard validation rules for Member States*

The agreed list of common validation rules that Member States are invited to apply is comprehensible without knowing details about waste statistics:

1. Comparison over time
   1. Waste generation, broken down by NACE activity
   2. Share of hazardous waste in total waste, broken down by NACE activity
   3. Waste treatment broken down by treatment operation (incineration, recovery, landfill etc.)
   4. Waste generation and waste treatment, by waste material
   5. Ratio of waste generation and treatment, by waste material
   6. Variation at detailed level:
      1. Waste generation: by waste material and NACE activity
      2. Waste treatment: by waste material and treatment operation
2. Ratio of waste generation and treatment, broken down by hazardous and non-hazardous waste
3. Waste treatment: implausible combination of waste material and treatment operation (for instance, incineration of mineral waste is unlikely).
4. Amounts of waste treated vs. capacities in waste treatment facilities (capacity should be ≥ amounts treated)

For these validation rules, deviation thresholds are suggested in the reference document using simple factors for the identification of outliers. We selected these validation rules according to the following criteria:

1. Validation rules should be relevant and applicable in all countries. Specifics like the economic structure of a country should not be taken into account.
2. The checks are limited to the datasets for the WStatR produced at national level. Cross-checks with other countries or related datasets within the country will remain Eurostat's task.
3. The validation rules do not require any reference variables, such as demographic or economic data (population, employees, value added etc.). Applying them to all datasets at once and consistently seems to be more efficient at Eurostat.
4. Checks of the file format and structural metadata (code lists) of the datasets do not require attention because EDAMIS web forms are used for data reporting and trans­mission. Member States automatically submit files that are correctly formatted and coded in SDMX-ML format using the web forms.

In terms of the validation levels described in figure 1, this means that level 2 checks are most appropriate and straightforward for a harmonised implementation in all Member States. However, outlier detection comparing countries is the only level 2 check that obviously remains at Eurostat. Level 0 and 1 checks should ideally be made obsolete by using efficient reporting standards and tools, while checks at levels 3 and higher seem to be better placed at Eurostat, at least in waste statistics. This means in practice that Eurostat continues checking WStatR datasets against related reporting obligations on particular waste streams (packaging waste, end-of-life vehicles, electronic waste, etc.) and against datasets e.g. collected by the OECD.

The forthcoming WStatR data collection, which is due by 30 June 2014 and which covers reference year 2012, will show the first results of the agreed division of labour on validation in waste statistics.

*2.5 Further activities*

Both the ESS.VIP Validation and its implementation in waste statistics are work in progress. We intend to continue the promising work because we can expect tangible efficiency gains from the ESS.VIP, both for the Member States and Eurostat.

As communication is essential to a common understanding of validation among statisticians at national and European level, Eurostat will offer an ESTP training course in September 2014 to present and discuss in detail the validation plan for waste statistics.

One observation at the workshop in September 2013 was that thresholds for outlier detection can be difficult to set in a meaningful way for all countries, waste materials or NACE activities because the absolute amount of waste generation or treatment has an effect on reasonable ranges of deviation. Put simply, a given percentage change of waste generation at the level of thousands of tonnes deserves less attention than at the level of millions of tonnes. Statistics Netherlands has suggested a logarithmic formula for the calculation of the maximum allowed change of waste amounts over time depending on the absolute value. Eurostat intends to examine further the use of such a formula.

The common validation rules for waste statistics are currently formulated in a clear but non-standard notation. Once VALS is finished, we intend to translate our validation rules into this syntax.

Finally, the generic software EDIT[[5]](#footnote-5) will be examined further as a possible common solution to automated validation of datasets in Member States. Eurostat has developed EDIT as a generic editing system for use by the ESS. Our first impression in waste statistics has been that EDIT is most suitable for level 0 and 1 checks; use for levels 2 and higher requires further investigation.

*2.6 Conclusion*

We can already observe now that starting vertical integration can bring about tangible benefits for statistical production in practice. Formulating standard validation rules and agreeing on their assignment to Member States and Eurostat have already been significant steps forward. The benefit of making validation more efficient in practice should be common to all official statistics regardless of the particular domain.

Further efforts to make validation in the ESS more efficient will gain from the outcome of the on-going ESS.VIP Validation and its results for horizontal integration. In particular, using a standard syntax such as VALS for describing validation rules and using it for automating processes in a future shared validation system should save a lot of resources both in Member States and at Eurostat.

1. https://webgate.ec.europa.eu/fpfis/mwikis/essvalidserv/index.php/Validation\_in\_the\_ESS [↑](#footnote-ref-1)
2. Regulation on waste statistics (EC) No. 2150/2002, amended by Commission Regulation (EU) No. 849/2010. <http://europa.eu/legislation_summaries/environment/waste_management/l28082_en.htm>. For more detail on waste statistics, visit <http://ec.europa.eu/eurostat/waste>. [↑](#footnote-ref-2)
3. <http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-RA-13-015> [↑](#footnote-ref-3)
4. ESS Standard for Quality Reports Structure, see http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/documents/ESQRS\_Structure.xls [↑](#footnote-ref-4)
5. http://ec.europa.eu/eurostat/edit/ [↑](#footnote-ref-5)