Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Implementing main types of International validation rules in national validation processes

Prepared by
Olav ten Bosch, Mark van der Loo, Statistics Netherlands, the Netherlands
Sónia Quaresma, Statistics Portugal, Portugal
Correspondence email: o.tenbosch@cbs.nl

I. Abstract

1. One of the major challenges in International data transmissions among NSIs and International organisations is reducing the number of resends and improving quality by improving validation mechanisms. From 2015 on multiple countries and Eurostat have been working on this. They developed validation principles, a handbook on validation, validation architectures, a standard validation report and they executed experiments implementing International validation rules in various statistical domains[1].

2. In 2018, Eurostat carefully examined the validation rules applied at their office and identified 21 'main types of validation rules' that were generic across domains. These rules, varying from simple range checks to more complex outlier detections, represent the majority of checks that are actually needed in today's practical International data validation exercises. The rules were generalized, documented in both natural language as well as in VTL (Validation and Transformation Language) and parametrized so that any domain can use them for International agreements.

3. Both Statistics Netherlands as well as Statistics Portugal implemented a considerable part of these main types of validation rules in their national validation languages. Instead of trying to automatically execute VTL code into their systems, they implemented the main types of rules in their national systems in a generic way and tested them across some statistical domains. The result are two base libraries implementing the same type of generic validation rules thus effectively standardizing International validation processes. The functionality is implemented in two different national languages: R-validate (by Statistics Netherlands) and SQL (by Statistics Portugal) and is available[2] for re-use by any other member state to build up International validation procedures in any statistical domain. Statistics Iceland and Statistics Hungary already started experiments using this approach.

4. This paper reports on these exercises. It describes the results and analyses the commonalities in these projects. It also motivates the potential of this approach which can be re-used in other contexts and other NSIs. Finally it sketches a future ESS validation landscape with a pool of generic data validation rules applicable in any statistical context in any statistical domain in any process phase.

[1] Overview validation results on CROS: https://ec.europa.eu/eurostat/cros/content/data-validation-overview_en
[2] https://github.com/SNStatComp/DomainValidationRules
https://github.com/SNStatComp/GenericValidationRules
https://github.com/SoniaQuaresma/MainTypeValidRules

II. Validation in the ESS

A. The problem

5. Just as with software, data contains errors. Just as with software we have to face the reality that in many cases we cannot easily guarantee its correctness. We can try to minimize errors by specifying the conditions (rules) that data have to satisfy, but even that sounds easier than it is in reality. Complex data chains – such as those executed in official statistics – build up rules which are scattered across different processing systems, executed in different process phases and which tend to grow organically over time. Data validation may sound easy in theory, it can be a difficult job in reality.

6. Data validation in the European Statistical System (ESS) is even more challenging. Data transfers among organisations – usually a national institute sending micro- or aggregated data to an international organisation – ideally has to meet the expectations of both data producer as well as data consumer. However, in many cases there is a misinterpretation or misunderstanding. Careful analysis of numerous international data transmissions by Eurostat over the past years have shown that in many cases data transfers had to be repeated over and over because of different data expectations on both sides. This phenomenon is also known as *'data ping pong'*. This happens not only between institutes, it happens in every data intense process also within national institutes.

B. Methodology

7. From 2015 on multiple European projects worked on the problem sketched in the previous section. One result is a handbook on methodology for data validation which describes levels of data validation, data validation procedures, a description of the business perspective, the data validation life cycle and some thinking on validation metrics, i.e. how to measure the success of a validation procedure. Since it is good to have a good understanding of the concept of data validation in this paper we quote the internally agreed definition from the handbook:

Data Validation is an activity verifying whether or not a combination of values is a member of a set of acceptable combinations.

For a detailed description of the other methodological concepts we refer to the handbook [1].

C. Validation Principles

8. In 2016 an international task force formulated a number of validation principles that are believed to comprise the most essential aspects in every validation step in any validation process. For completeness and understanding of the rest of this paper we repeat them briefly with an informal one-liner description:

(a) *The sooner, the better*: do your validation as early as possible in your data process
(b) *Trust but verify*: data validation builds on trust and validation both sides
(c) *Well-documented and appropriately communicated validation rules*: rules have to be agreed upon on both sides
(d) *Well-documented and appropriately communicated validation errors*: there is no point in saying that the data is wrong, one has to specify what is wrong, where in the data and by which rule
(e) *Comply or explain*: once volition rules are agreed upon, then obey them or explain in case there is really an exception
(f) *Good enough is the new perfect*: there is no point in trying to be complete in specifying validation rules, the trick is to use just enough rules to guarantee sufficient data quality for the target output.

For a more extensive explanation of these principles we refer to [2].

D. Alignment of international and national validation processes

9. The process of data validation can be seen as part of the wider scope of statistical data editing (SDE). For this wider scope the UNECE created a generic process model, the Generic Statistical Data Editing Model (GSDEM) [3]. Data validation is part of the 'Review' business function: functions that examine the data to identify potential problems. For the subject of this paper it is important to realize that international data validation processes have an effect on the statistical processes within National institutes. International rules have to be checked at the output of National processes before sending, but, in-line with the principle of 'the sooner the better' it is better to check them even earlier in national statistical process chains. Stated differently, international validation rules 'travel' from output to input through international process chains.

10. Observing the fact that international rules should ideally somehow be implemented in national processes, the question arises how to do this. Keep in mind that NSIs have many different processing systems, each with different characteristics with respect to architecture, IT and processes. One view on this is to develop a formal language in which the validation rules can be expressed so that every statistical authority can parse the rules and automatically execute them in their national context. This is the idea behind the development of the Validation and Transformation Language (VTL) [4] by the SDMX consortium [5]. One other approach is to identify a set of generic validation rules being used throughout the ESS and express the domain specific rules in terms of the generic ones. This is the approach taken by Eurostat in the study which identifies 21 'main types of rules' in the ESS [6]. These main types of validation rules are expressed in both natural language as well as in VTL.

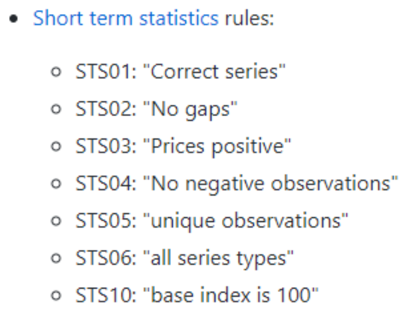
11. The two approaches sketched above do not necessarily conflict, they may go hand in hand. We will come back to this later. First, in the next two chapters, we dive into two pilot projects in two countries that focussed on implementing the second approach, implementing international rules in a national context using the main types of rules as much as possible.

III. Pilot1: implementing international rules in the Netherlands

12. Based on an analysis of sufficiently-documented international validation rules and interviews with various national experts within Statistics Netherlands two statistical domains were chosen to start piloting the implementation of international rules: the short term statistics (STS) and National Accounts (NA). For the STS there is an internationally agreed reference document [7] containing 11 internationally agreed validation rules, expressed in natural language, which – at that time - were not fully checked nationally. For NA there are many interesting and well-documented rules in the SNA [8] which could serve as a good example. Furthermore it was decided to work on a generic implementation in the national context of the main types of validation rules as identified by Eurostat. We describe the results of these activities in the next two sections.

A. Domain specific rules

13. Reviewing the production process for STS at Statistics Netherlands it became clear that the STS is in fact being composed from multiple internal data providers each delivering a specific part of the data to be reported. This makes the use case even more interesting. This means that the international rules can be checked just before the data is reported internationally but also during the preceding internal data streams. Doing the latter is in fact completely in-line with the number one validation principle "the sooner the better".



- Short term statistics rules:
  - STS01: "Correct series"
  - STS02: "No gaps"
  - STS03: "Prices positive"
  - STS04: "No negative observations"
  - STS05: "unique observations"
  - STS06: "all series types"
  - STS10: "base index is 100"

Figure 1: STS validation rules

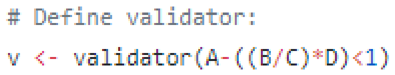
14. Figure 1 shows the STS rules that were successfully implemented in the pilot. A few observations can be made from the implementation activities of these rules and also from the inability to implement the other 4 rules:

(a) The rules in itself were not too complicated. All of the rules could be implemented in a single line of R-validate syntax.
(b) Some of the rules, such as Rule STS02 “no gaps” are very generic by nature. These were implemented using a generic function (RTS) from the Eurostat main types of rules. This proofs the validity of implementing generic rules and applying them in domain specific context.
(c) There is a relationship of the rule with the metadata as specified in the internationally agreed SDMX data structure descriptions (DSD). A flexible implementation of such validation rules can profit of access to this metadata. An example is the check on positive prices for which the SDMX metadata contains the definition of the price variables. We conclude that such functionality might better be implemented in a generic way than in a specific implementation.
(d) Some of the rules expressed in natural language could not be implemented in this pilot because of lack of preciseness. The textual description lacked an exact description or reference to the exact method to apply calendar effects and seasonally adjustments. Since these types of checks are very common in the ESS these could better be implemented in a generic module based on a generic definition of validation rules (see later).
(e) Some of the rules in natural language were not implemented because the project had difficulties interpreting the exact meaning. This could be solved by either using a more exact language for specifying validation rules or by improving the textual description and adding sufficiently documented examples.

15. The system of nation accounts (SNA) contains a large number of validation checks that should be executed before transfer. Some of these checks were documented in the new generic validation language VTL 2.0 by Eurostat. Since these VTL rules were, at that stage, still work in progress the project decided to implement one rule that was thought to be important as well as a good test case: the “chain linking formula” as described in the ESA 2010 Handbook on Data Validation.
The result was that the project did implement one of the National Account rules, mainly to see if it was doable and to get some experience applying a VTL 2.0 rule to the National Accounts data using the National R-validate syntax. It was concluded that if the rules are expressed in the right way, whether it is VTL or another language, it is certainly possible to implement them in the open source validation tool set.

16. We made the following observations from this effort:

(a) The majority of the code implementing this rule involves selecting the right slice of data from the NA database. This was not straightforward, it involves filtering on multiple variables and transforming it into a the right format.
(b) The actual implementation of the chain linking formula rule is only one line of code, shown below in the R-validate syntax:



```
# Define validator:
v <- validator(A-((B/C)*D)<1)
```

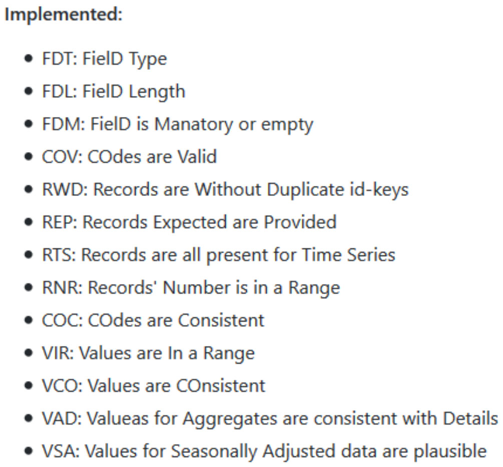
The implementations of both the STS rules as well the NA rule in the R-validate language have been published on GitHub at:

https://github.com/SNStatComp/DomainValidationRules

B. Generic rules

17. The document describing the main types of validation rules provided by Eurostat was analysed. A choice was made which rules to implement. This choice included all of the “Basic intra-file check rules” and almost all of the “Check intra or inter files rules”. This choice aligns well with the choice that was made by fellow Grant partner Portugal for their implementation in SQL. This proves that it is indeed possible to implement the generic rules into different national languages.

18. The rules were implemented in a generic way, with the aim to apply them in multiple domains, and possibly multiple statistical offices. The data of the examples provided in the document were extracted and added to the GitHub repository used to implement the R-validate implementation. Automatic tests were defined to execute the examples from the document upon each code change. The implementation of the rules has been packed into a R-package, published on GitHub. Documentation has been written in the R-style which has the advantage that it is available in context-sensitive help in R and/or RStudio. The names of the main types of rules (FDT, FDL etc.) were used as names for the generic R-functions. Figure 2 shows the main types of rules that were implemented:



Implemented:

- FDT: FielD Type
- FDL: FielD Length
- FDM: FielD is Manatory or empty
- COV: COdes are Valid
- RWD: Records are Without Duplicate id-keys
- REP: Records Expected are Provided
- RTS: Records are all present for Time Series
- RNR: Records' Number is in a Range
- COC: COdes are Consistent
- VIR: Values are In a Range
- VCO: Values are COnsistent
- VAD: Valueas for Aggregates are consistent with Details
- VSA: Values for Seasonally Adjusted data are plausible

Figure 2: Main types of validation rules implemented

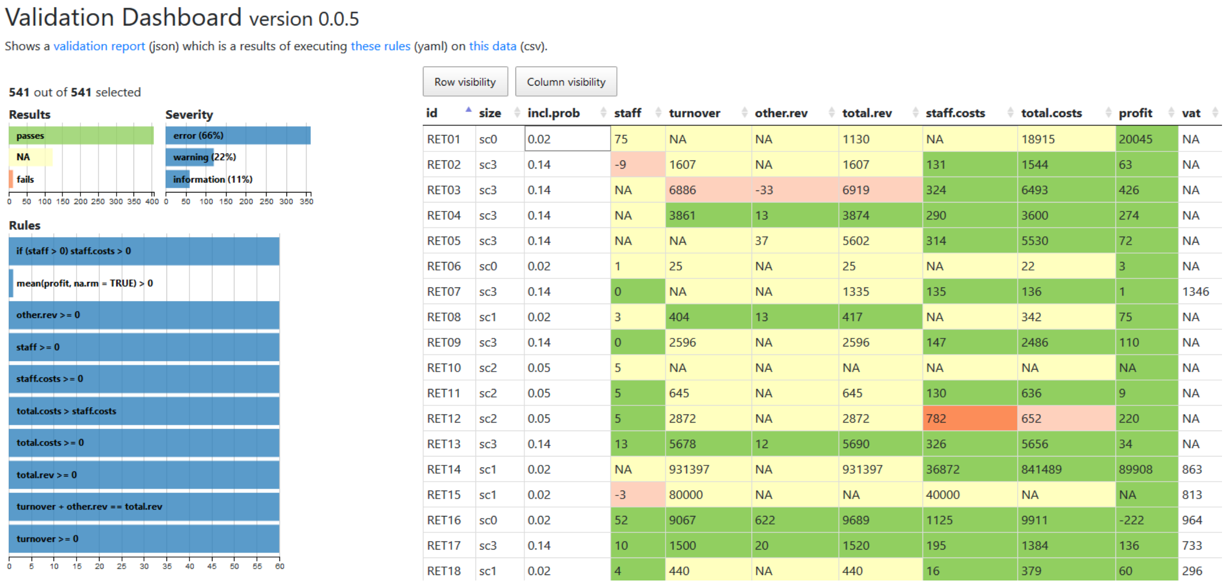
19. The implementation of the generic rules in the R-validate syntax has been packaged in an R-package. The README.md found at the following link contains instructions how to install this package locally so that everyone can use it. For further information on its use we refer to the R-package documentation which is installed together with the R-package. The results can be found at:

https://github.com/SNStatComp/GenericValidationRules

C. Validation dashboard

20. The validation dashboard, presented in an earlier papers [9, 10], which is meant to easily analyse validation results was redesigned. One extension in functionality is the facility to view the validation results graphically in a data grid (datatable) displaying the data. There is an on-line demo page [11]. The links to dashboard7 and dashboard8 show respectively examples on synthetic population data and a perturbed dataset reflecting some synthetic SBS data.

21. Figure 3 shows a screenshot of the dashboard displaying the synthetic SBS data and the validation errors. Validation results are coloured green (passes), yellow (NA) or red (failures). Selecting one or more result types, severities or rules, acts as a filter on the selected validation events and the colouring in the data grid will be updated accordingly. Selecting a specific cell in the data grid also acts as a filter. The bar charts on the left will be updated accordingly.



Validation Dashboard version 0.0.5

Shows a validation report (json) which is a results of executing these rules (yaml) on this data (csv).

541 out of **541** selected

id	size	incl.prob	staff	turnover	other.rev	total.rev	staff.costs	total.costs	profit	vat
RET01	sc0	0.02	75	NA	NA	1130	NA	18915	20045	NA
RET02	sc3	0.14	-9	1607	NA	1607	131	1544	63	NA
RET03	sc3	0.14	NA	6886	-33	6919	324	6493	426	NA
RET04	sc3	0.14	NA	3861	13	3874	290	3600	274	NA
RET05	sc3	0.14	NA	NA	37	5602	314	5530	72	NA
RET06	sc0	0.02	1	25	NA	25	NA	22	3	NA
RET07	sc3	0.14	0	NA	NA	1335	135	136	1	1346
RET08	sc1	0.02	3	404	13	417	NA	342	75	NA
RET09	sc3	0.14	0	2596	NA	2596	147	2486	110	NA
RET10	sc2	0.05	5	NA	NA	NA	NA	NA	NA	NA
RET11	sc2	0.05	5	645	NA	645	130	636	9	NA
RET12	sc2	0.05	5	2872	NA	2872	782	652	220	NA
RET13	sc3	0.14	13	5678	12	5690	326	5656	34	NA
RET14	sc1	0.02	NA	931397	NA	931397	36872	841489	89908	863
RET15	sc1	0.02	-3	80000	NA	NA	40000	NA	NA	813
RET16	sc0	0.02	52	9067	622	9689	1125	9911	-222	964
RET17	sc3	0.14	10	1500	20	1520	195	1384	136	733
RET18	sc1	0.02	4	440	NA	440	16	379	60	296

Figure 3: Validation dashboard showing validation statistics and data coloured by validation results.

D. Putting it all together

22. Now that we described the pilots implementing generic validation rules, domain specific validation rules and a generic validation dashboard we can put together the individual pieces to show the result. Figure 4 shows the three phases of data validation, which comprises the definition and maintenance of the rules, the process of validating data and the analysis phase. In the pilot for STS some of the domain specific rules were implemented in terms of the parametrized generic rules. The software being used is part of the official statistics open source ecosystem which is listed on the awesome list of official statistics software [12].

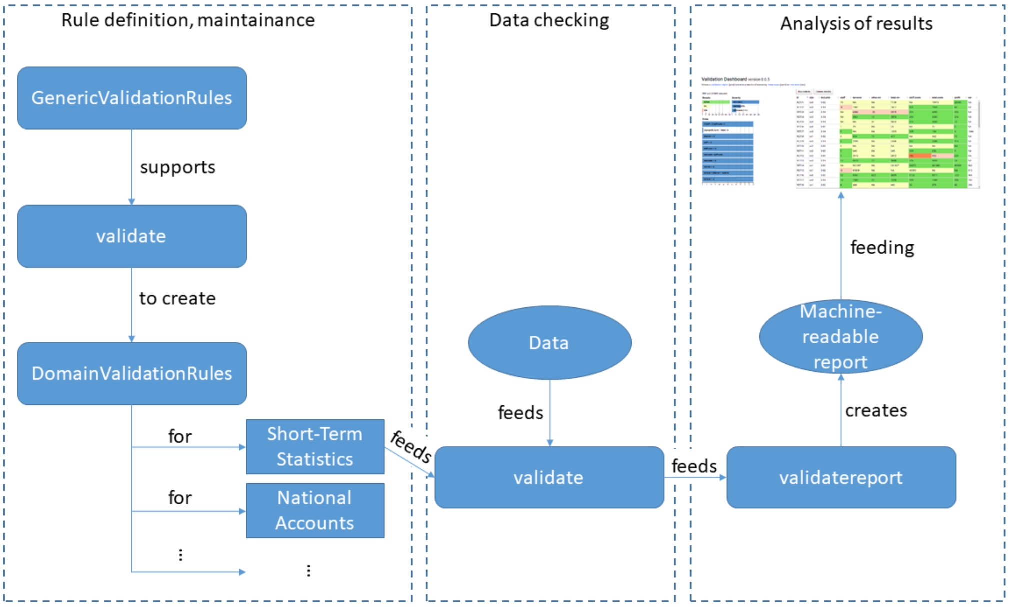


Figure 4: Data validation workflow

IV. Pilot2: implementing international rules in Portugal

23. In Portugal a similar project was executed implementing international validation rules in national statistics. This project was called the Hybrid Validation Implementation Project (HyVImp). Although the domain chosen differs from the pilots executed in the Netherlands, the rules that were implemented were partly the same, making it an ideal case to compare national initiatives. In Portugal the project focus was on the ANIMAL domain.

24. Statistics Portugal performed a manual translation of the domain specific rules and the main types of validation rules into the language of the national system. The starting point was the national Statistical Data Warehouse (SDW), an Oracle database, where all national data is already integrated. This database also includes all the national validation rules and could therefore also easily be extended to include the Eurostat's set of rules. A transformation was done from the VTL specification of the rules into SQL. A template was developed to offer parametrized code implementing the generic rules in specific domains. To facilitate implementation the script automatically generates an execution log on execution.

25. This approach has the following advantages:

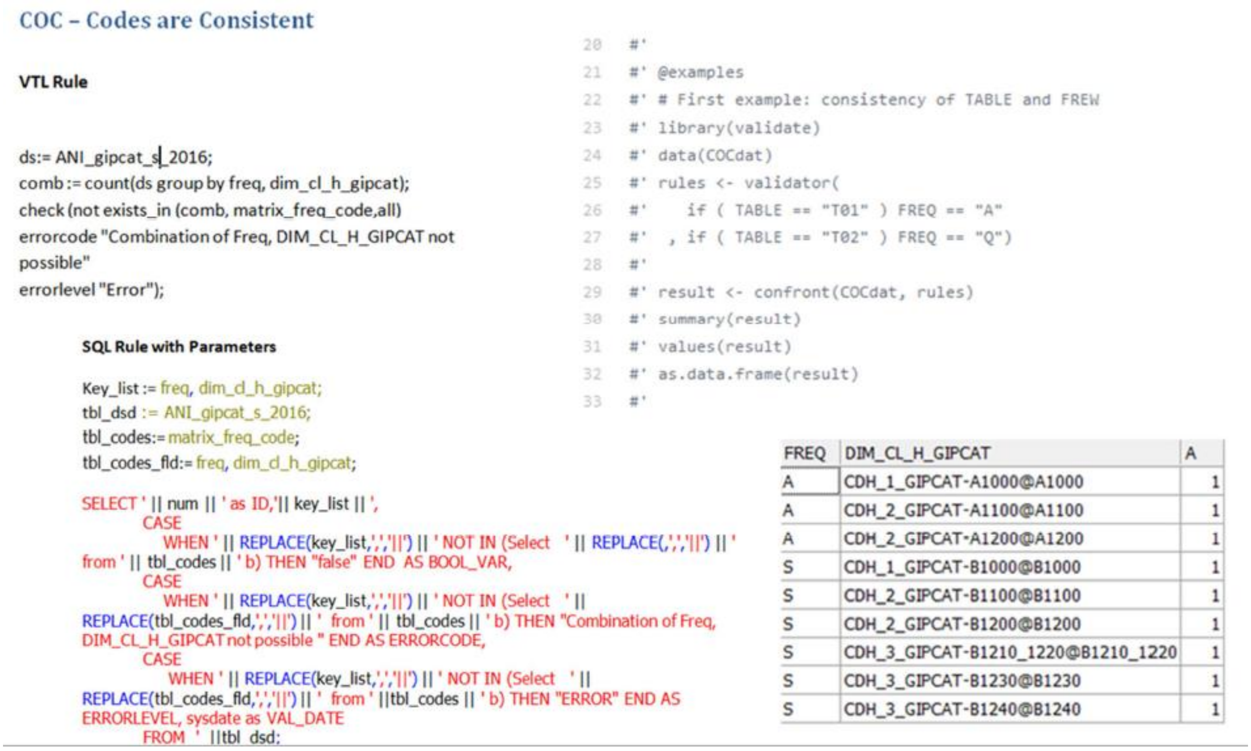
(a) It facilitates the construction of specific rules in any domain. The parametrized validation rules can be implemented quite easily in other domains and the domain knowledge is as much as possible encapsulated in the parameters. Domain specialists don't necessarily need to know the deeper implementation details of the rules. This increases the autonomy of the domain specialists and at the same time lowers the burden on the IT department.

(b) Central maintenance of validation rules in the SDW allows methodologists to monitor and control in a systematic and organized way not only which specific rules are being used in any domain, but also which types of rules are applied per domain. This also makes it easier to harmonize the validation process across domains and evaluate the rules being used.

(c) Because most data is already stored in the central data warehouse, it becomes easy to automatically generate files in the SDMX format conforming international validation requirements. Mechanisms developed for ANIMAL can be re-used in other domain. The next project scheduled is on the demographics area, in particular for migrations.

26. Figure 5 shows an example of the implementation of one of the main types of rules: COC- codes are consistent. Both the VTL code as well as the parametrized SQL code are shown. The full implementations of the international rules both the source code as well as the choices made and the assumptions regarding the database structure, have been published on Github at the following address:

https://github.com/SoniaQuaresma/MainTypeValidRules



COC – Codes are Consistent

VTL Rule

```
ds:= ANI_gipcat_s_2016;
comb := count(ds group by freq, dim_cl_h_gipcat);
check (not exists_in (comb, matrix_freq_code,all)
errorcode "Combination of Freq, DIM_CL_H_GIPCAT not
possible"
errorlevel "Error");
```

SQL Rule with Parameters

```
Key_list := freq, dim_cl_h_gipcat;
tbl_dsd := ANI_gipcat_s_2016;
tbl_codes:= matrix_freq_code;
tbl_codes_fld:= freq, dim_cl_h_gipcat;

SELECT ' || num || ' as ID,'|| key_list || ',
        CASE
            WHEN ' || REPLACE(key_list,',','||') || ' NOT IN (Select   ' || REPLACE(,',','||') || '
from ' || tbl_codes || ' b) THEN "false" END  AS BOOL_VAR,
        CASE
            WHEN ' || REPLACE(key_list,',','||') || ' NOT IN (Select   ' ||
REPLACE(tbl_codes_fld,',','||') || '  from ' || tbl_codes || ' b) THEN "Combination of Freq,
DIM_CL_H_GIPCAT not possible " END AS ERRORCODE,
        CASE
            WHEN ' || REPLACE(key_list,',','||') || ' NOT IN (Select   ' ||
REPLACE(tbl_codes_fld,',','||') || '  from ' ||tbl_codes || ' b) THEN "ERROR" END AS
ERRORLEVEL, sysdate as VAL_DATE
        FROM  '  ||tbl_dsd;
```

```
#'
#' @examples
#' # First example: consistency of TABLE and FREW
#' library(validate)
#' data(COCdat)
#' rules <- validator(
#'     if ( TABLE == "T01" ) FREQ == "A"
#'   , if ( TABLE == "T02" ) FREQ == "Q")
#'
#' result <- confront(COCdat, rules)
#' summary(result)
#' values(result)
#' as.data.frame(result)
#'
```

FREQ	DIM_CL_H_GIPCAT	A
A	CDH_1_GIPCAT-A1000@A1000	1
A	CDH_2_GIPCAT-A1100@A1100	1
A	CDH_2_GIPCAT-A1200@A1200	1
S	CDH_1_GIPCAT-B1000@B1000	1
S	CDH_2_GIPCAT-B1100@B1100	1
S	CDH_2_GIPCAT-B1200@B1200	1
S	CDH_3_GIPCAT-B1210_1220@B1210_1220	1
S	CDH_3_GIPCAT-B1230@B1230	1
S	CDH_3_GIPCAT-B1240@B1240	1

Figure 5: Implementation of COC rule in Statistical Data Warehouse in Portugal

V. Discussion and conclusion

27. Both Statistics Netherlands as well as Statistics Portugal experimented with the implementation of main types of international validation rules into national languages. The national language is different, but the concept is the same: implementing main types of rules in parametrized functions or templates so that they can easily be applied in domain specific contexts.

28. In section II.D we noted that there are in principle (at least) two approaches to implementing international data validation: (1) developing software (a VTL parser and VTL execution environment) that automatically executes international rules expressed in VTL on national data or (2) transforming the recognized main types of rules in generic functions / templates and apply them in domain specific contexts as in the pilots described. Although the future validation landscape might be a combination of both, we claim that the projects presented in this paper at least proof that the second scenario is also a viable option for validation in the ESS. One necessary pre-condition is that international rules are as much as possible expressed in terms of these main types of rules of future versions. If this can be achieved it might influence the choices [13] NSIs have implementing validation in national context

29. The main types of rules developed by Eurostat were identified from a careful examination of the rules being used in multiple statistical domains over the past years. Recognizing the potential of the second scenario, we think it would be very useful to build on these initial version of the main types of rules, not only by adding new use cases from practice, but also from a more formal point of view. The challenge would be to develop a minimum set of parametrized generic validation rules that would cover most or all of the validation needs in the ESS. In an ideal future the statistical community could maintain implemented version of these ‘extended main types of rules’ in the implementation languages needed. We hope to work on such approach in future research.

References

[1] Methodology for data validation 2.0. Revised Edition 2018. https://ec.europa.eu/eurostat/cros/system/files/ess_handbook_-_methodology_for_data_validation_v2.0_-_rev2018_0.pdf
[2] Principles for data validation, https://ec.europa.eu/eurostat/cros/content/principles_en
[3] Generic Statistical Data Editing Model, version 2, 2019, https://statswiki.unece.org/display/sde/GSDEM
[4] VTL 2.0 (Validation and Transformation Language), July 2018, https://sdmx.org/?page_id=5096
[5] SDMX: Statistical Data and Metadata eXchange, www.sdmx.org
[6] V. Tronet (2018), Main types of validation rules for ESS data (version 1.0.3). Eurostat Working document.
[7] SDMX for Short-Term Business Statistics Guidelines (STS). https://circabc.europa.eu/sd/a/adfce7f9-49ff-47cf-8a47-4224011a8d48/SDMX%20for%20STS_guidelines_170201.pdf
[8] ESA 2010 - Handbook on Data Validation, updated 2019, https://webgate.ec.europa.eu/fpfis/wikis/display/ESRNA/ESA+2010+-+Handbook+on+Data+Validation
[9] A generic validation report for the ESS, O. ten Bosch, M. van der loo, UNECE Workshop on Statistical Data Editing, Sep. 2018, Neuchatel, https://www.unece.org/index.php?id=47802
[10] O. ten Bosch and Van der Loo, M. (2019). A generic Shiny/JS dashboard for data validation results. Use of R in Official Statistics (uRos2019, Bucharest).
[11] https://github.com/data-cleaning/ValidatReport
[12] Awesome list of official statistics software: http://www.awesomeofficialstatistics.org . Official statistics software is ‘awesome’ when it is (1) Free, Open Source, and available for download, and (2) confirmed to be used in the production of official statistics by at least one institute. Software that facilitates access to official statistics is accepted as well, as long as it conforms to (1).
[13] Validation Scenarios for member states: https://ec.europa.eu/eurostat/cros/content/scenarios-member-states_en