A generic validation report for the ESS

Statistics Netherlands

Olav ten Bosch and Mark van der Loo

19-09-2018

Contents

- Why data validation?
- ESSnet projects and validation principles
- Design of the validation report
- Machine and human readable versions
- Implementation in R
- Wrap up

Why data validation? (1)

Data ping pong:

- ***Resending*** data files again and again
- The record seems to be 21 re-transmissions
- ***Multiple*** NSI's and ***multiple*** domains
- Has to be solved ***together***
- ***Validate*** data before sending

Why data validation? (2)

∑V21110(2I) over all size classes<=V15110(2A)
Total investment in equipment and plant for pollution control, and special anti-pollution accessories (mainly end-of-pipe equipment)over all size classes should not exceed total investments

6	V16110>=V11110	Warning
	Each enterprise should employ at least one person. Enterprises without any persons employed are possible however	

SBS

FSS

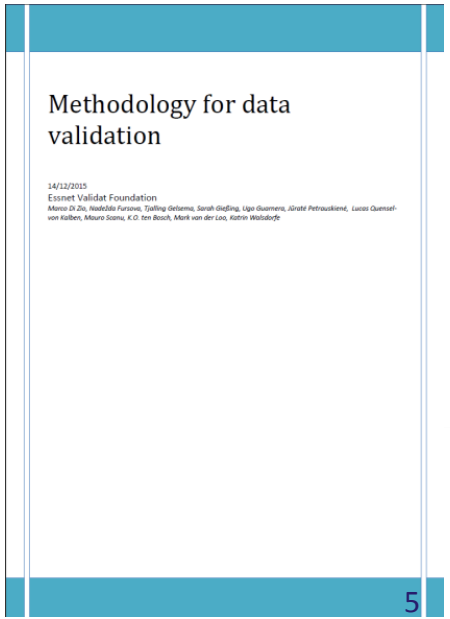
Ruleno	Condition
1	(E_1_2$WorkCodeM='24' and E_1_2$AWU>0 and E_1_2$AWU<0.25) or (E_1_2$WorkCodeM='49' and E_1_2$AWU>=0.25 and E_1_2$AWU < 0.5) or (E_1_2$WorkCodeM='74' and E_1_2$AWU>=0.5 and E_1_2$AWU<0.75) or (E_1_2$WorkCodeM='99' and E_1_2$AWU>=0.75 and E_1_2$AWU<1) or (E_1_2$AWU=1 and E_1_2$WorkCodeM='100')
2	if A_2$holdingtype in ('1','2a','2b','5') then E_1_2$AWU <= E_1_3$AWU end

Two ESSnets projects

ESSnet Validat Foundation 2015-2016 (DE, IT, LT, NL, ESTAT)

ESSnet Validat Integration, 2017 (DE, NL, LT, SW, PL, PT)

- Handbook on validation
- A study on VTL 1.0
- PoC with 3 national validation languages
- Business architecture scenario's
- Generic validation report



Validation principles

1. The sooner, the better
2. Trust, but verify
3. Well-documented and appropriately communicated validation rules
4. Well-documented and appropriately communicated validation errors
5. Comply or Explain
6. Good enough is the new perfect

Demands (1)

Principle 4: “Well-documented and appropriately communicated validation errors”

Principle 1: “The sooner, the better”

We need a standard validation report that can be used in:

- ***Every*** statistical ***domain***
- ***Every*** statistical validation ***tool***
- For ***ESS*** date ping pong as well as ***within NSI’s***
- For ***micro***data as well as ***aggregated*** data

Machine readable (JSON) and ***human readable*** version

Demands (2)

Identifiable:

- Every validation result must be fully identifiable in the business context

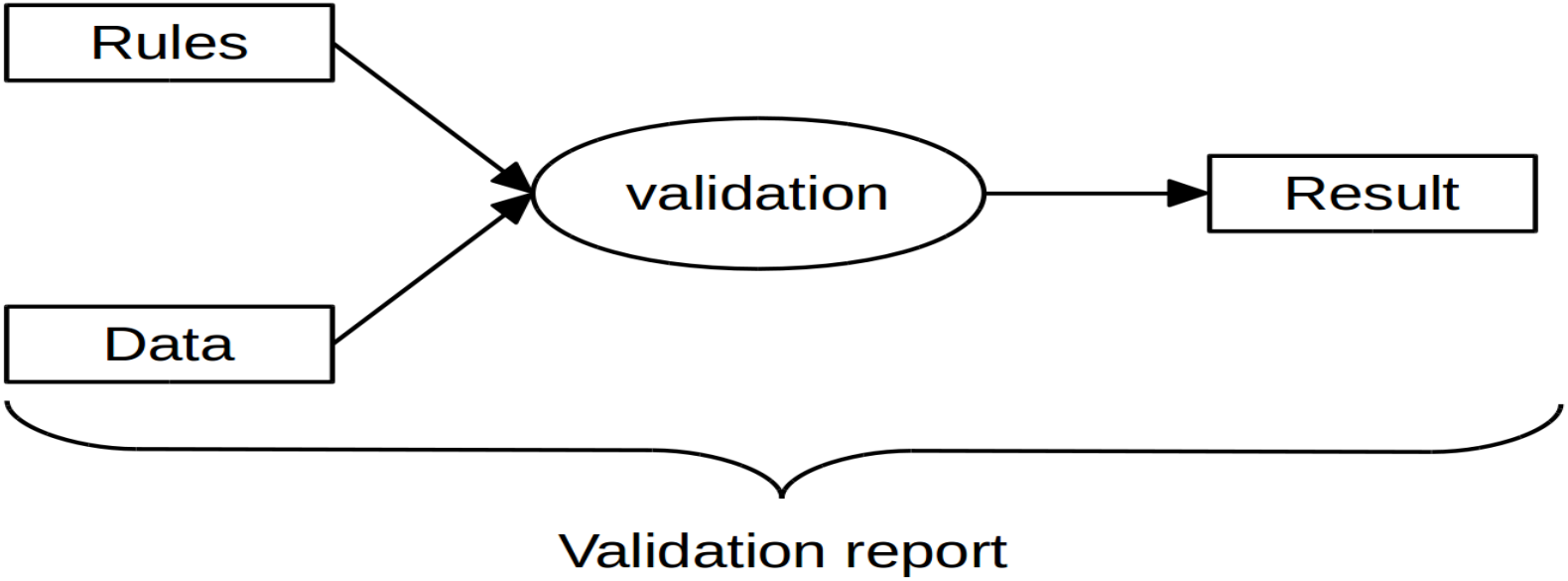
Composable:

- Two reports can be combined into a new report which is still a validation report (elements fully identifiable)

Aggregable:

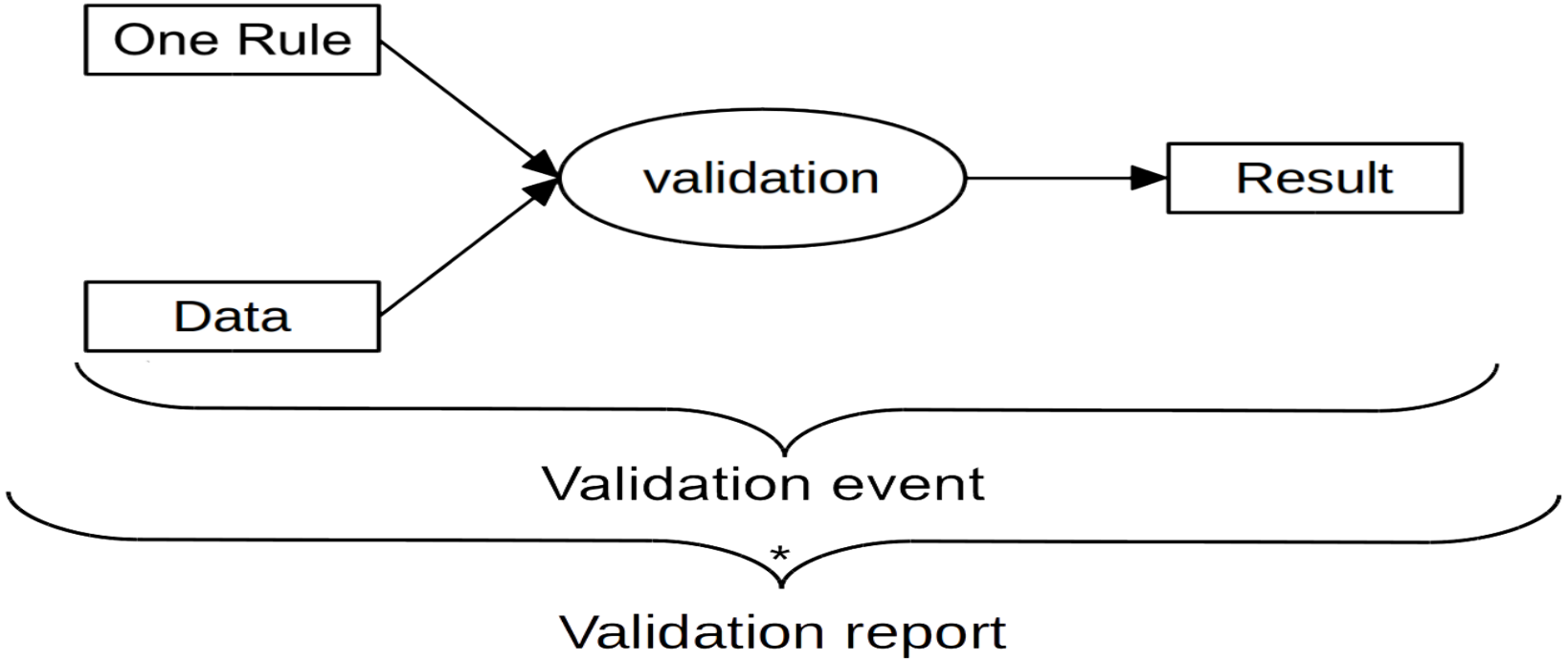
- Validation report can contain aggregates of validation events

Conceptual workflow (1)



Standard ESS validation report (pdf)

Conceptual workflow (2)



Logical information model

- **Event**
  - timestamp
  - actor (who/what)
  - *agent*, *trigger*
- **Rule**
  - Language
  - Expression
  - Severity (information/warning/error)
  - *Description*
- **Data**
  - $U\tau uX$ (population, measurement, population element, variable)
  - *Description*
- **Value** $(0, 1, \mathrm{NA})$

Machine readable example

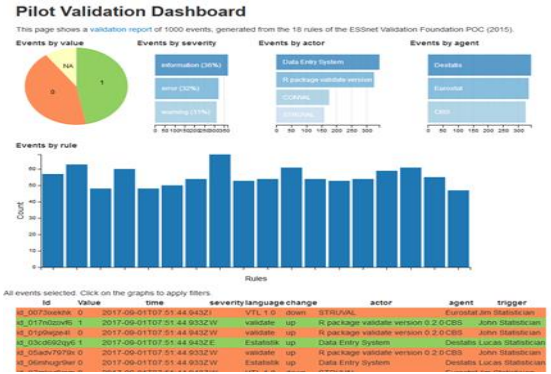

```
{
  "event": {
    "time": "20170518T105055+02",
    "actor": "R 3.4.0",
    "agent": null,
    "trigger": null
  },
  "rule": {
    "language": "R pkg validate 0.1.7",
    "expression": "income >= 0",
    "severity": "error",
    "description": "total income must be non-negative"
  },
  "data": [
    "Dutch inhabitants",
    "Household survey 2017",
    "8237193679",
    "Household Income"
  ],
  "value": "1"
}
```

From machine to human readable

Machine


```
[
  {
    "rule": {
      "expression": "check(DS.hours_worked between 1 and 80)",
      "severity": "warning"
    },
    "event": {
      "time": "2017-09-01T07:51:44.933Z",
      "actor": "Eurostat"
    },
    "data": {},
    "value": "0" // failed
  },
  {
    "rule": {
      "expression": "cost + profit == turnover",
      "severity": "error"
    },
    "event": {
      "time": "2017-09-01T07:51:46.933Z",
      "actor": "Eurostat"
    },
    "data": {},
    "value": "1" // passed
  },
  ...
  ...
  ...
]
```


Human readable



- View
- Filter
- Aggregate
- OSS: used in NL and Poland

[DEMO]

markdown

html **pdf** **word**

Implementation in R

R package ***validate***:

- implements concepts of the ESS *handbook* on validation
- On CRAN and awesome list 

R package ***validatetools***:

- Functions for finding *redundancies* or *contradictions*
- On CRAN and awesome list 

R package ***validatereport***:

- implements the *validation report standard*
- GH: https://github.com/data-cleaning/validatereport
- Improvements in 2018/19 towards CRAN and 

Wrap up

- ***Data ping pong*** in the ESS needs to be solved
- We developed a ***generic*** validation report:
  - For ***every*** statistical ***domain***
  - For ***every*** statistical validation ***tool***
  - For ***ESS*** date ping pong as well as ***within NSI's***
  - For ***micro***data as well as ***aggregated*** data
- ***Machine*** readable ad well as ***human*** readable
- ***Extensible*** for use in national systems
- Has been implemented in software from ***NL*** and ***Poland***. ESTAT studies its applicability in ESS tools.

Questions, ideas, suggestions

?

Olav ten Bosch
Mark van der Loo
obos@cbs.nl
mplo@cbs.nl

Curated list of software for official statistics

awesome

www.awesomeofficialstatistics.org