Can we Quantify the Quality of the Data?

Stefano Cappè

Data Management Team Leader
Evidence Management Unit
European Food Safety Authority

Eurachem Dublin, 14-15 May 2018

efsa
European Food Safety Authority

Summary

- Background
- Engagement from the start
- Quality of data
- Quantification of data quality
- Visualising data quality
- Governance and collaboration with Member States
- Conclusions

Background: Founding regulation

Regulation (Ec) No 178/2002 – Article 33

The MSs shall take the necessary measures to enable the data they collect in the fields of EFSA be transmitted to the Authority

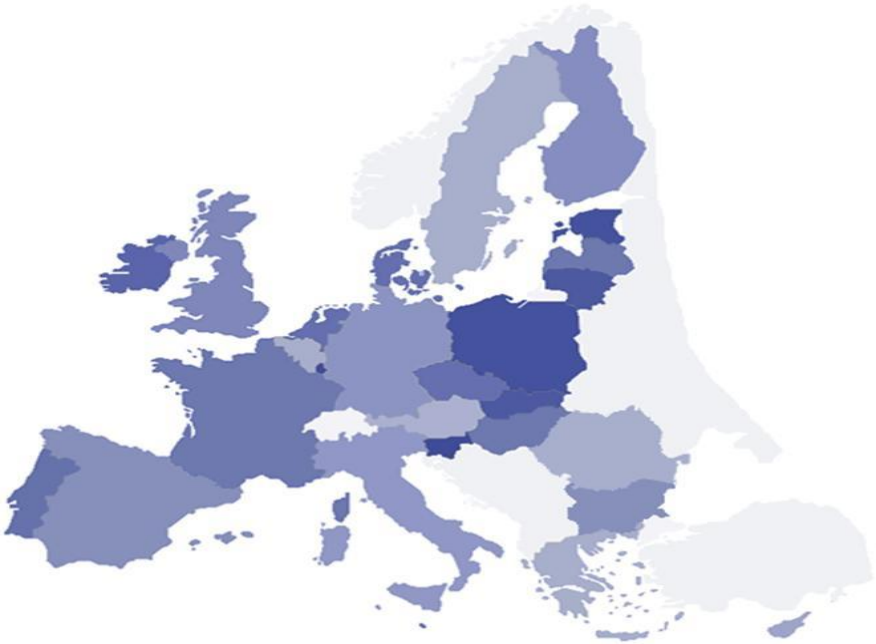
Data Collection

EFSA forward to Member States and EC appropriate recommendations which might improve the technical comparability of the data

Data harmonisation

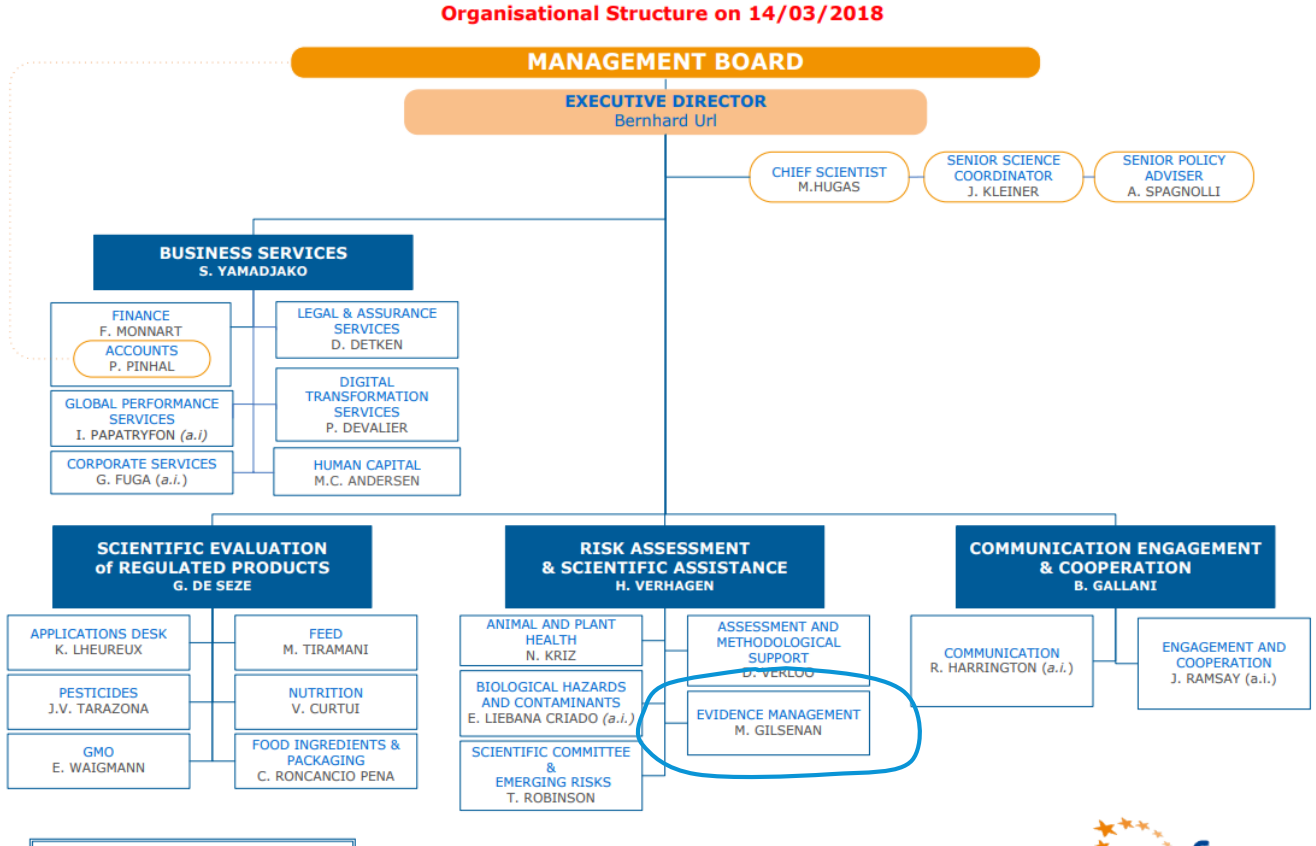
Background: Data collection

EFSA has no laboratories



The Laboratories in the Member States are EFSA's Laboratories

Background: Evidence management unit



Summary

- Background
- Engagement from the start
- Quality of data
- Quantification of data quality
- Visualising data quality
- Governance and collaboration with Member States
- Conclusions

A priority from the start: Data harmonisation initiatives

harmonization

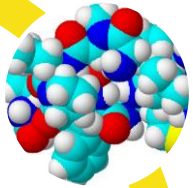
2010 - Standard Sample Description ver. 1

2013 - Standard Sample Description ver. 2

Based on XML (eXtensible Markup Language)
Data format specified with XML schemas

A priority from the start: Standard terminologies

Around 120 catalogues



Parameters (PARAM)

Matrices(FoodEx2)

Analytical methods(ANLYMD)

Around 200,000 term identities

Standardised concepts

Unique identities

Language independent

Reusable

Around 350,000 connected elements (e.g. CAS numbers, species latin names)

A priority from the start: Standardised validation rules

- Validation rules are implemented in XML language

businessRule Set

name

businessRuleCode

busine

scien

techr

Business Rules conforming with Guidance on Data Exchange version 2 (WF2)

Business Rules set

Name

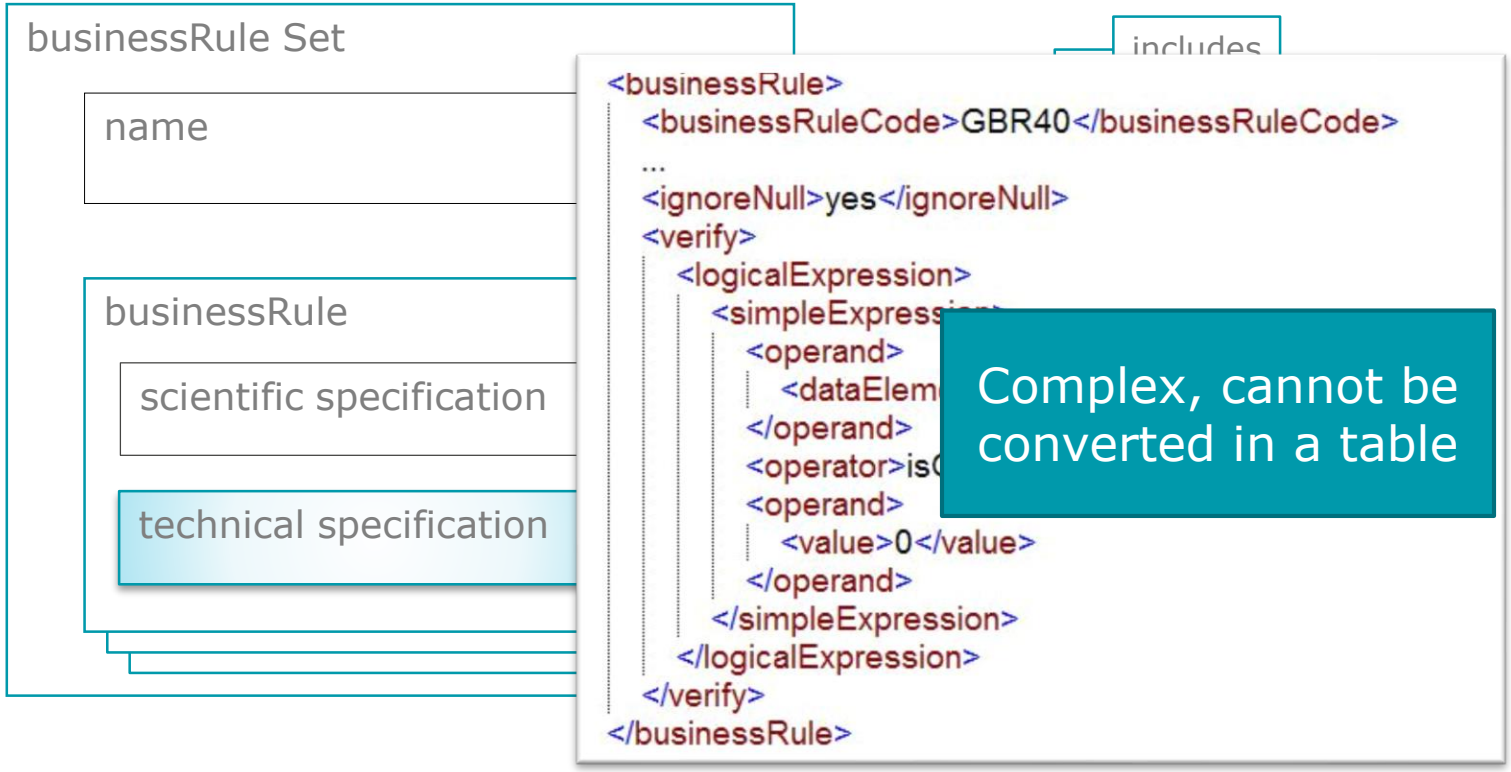
General business rules for SSD2

Business Rules List

Business Rule Code	Description	Error Message	Type of error	Status	Last Update	Checked Variables
GBR38	The value in 'Result LOD' (resLOD) must be greater than '0';	resLOD is not greater than '0';	error	active	2014-08-08	resLOD
GBR39	If the value in the data element 'Type of result' (resType) is 'Non Quantified Value (below LOQ)' (LOQ), then a value must be reported in the data element 'Result LOQ' (resLOQ);	resLOQ is missing, though resType is 'Non Quantified Value (below LOQ)' (LOQ);	error	active	2014-08-08	resLOQ resType
GBR40	The value in 'Result LOQ' (resLOQ) must be greater than 0;	resLOQ is not greater than 0;	error	active	2014-08-08	resLOQ
GBR41	If the value in the data element 'Type of result' (resType) is 'Value below CCalpha (below CCα)' (CCA), then a value must be reported in the data element 'CC alpha' (CCalpha);	CCalpha is missing, though resType is 'Value below CCalpha (below CCα)' (CCA);	error	active	2014-08-08	CCalpha resType
GBR42	The value in 'CC alpha' (CCalpha) must be less than the value in 'CC beta' (CCbeta);	CCalpha is not less than CCbeta;	error	active	2014-08-08	CCalpha CCbeta
GBR43	The value in 'CC alpha' (CCalpha) must be greater than '0';	CCalpha is not greater than '0';	error	active	2014-08-08	CCalpha
GBR44	If the value in the data element 'Type of result' (resType) is 'Value below CCbeta (below CCβ)' (CCB), then a value must be reported in the data element 'CC beta' (CCbeta);	CCbeta is missing, though resType is 'Value below CCbeta (below CCβ)' (CCB);	error	active	2014-08-08	CCbeta resType
GBR45	The value in 'CC beta' (CCbeta) must be greater than '0';	CCbeta is not greater than '0';	error	active	2014-08-08	CCbeta

A priority from the start: Standardised validation rules

- Business rules are implemented in XML language



A priority from the start: Acknowledgement message

Provide an automatic feedback to data providers:

Standard Sample Description Acknowledgment

Header

Type	dcfmsg
Version	1.0
Code	Example1.xls
Receiver's Code	EFSA
Sender's Code	EFSA
Sent date	2011-11-30T12:16:37.505

Message

Message Receive Date	2011-11-30T12:18:48.146+01:00
Message Ack Date	2011-11-30T12:18:48.146+01:00
Transmission Ack Code	02
Sender's Transaction Code	Example1.xls
Receiver's Transaction Code	7081
Data Collection Code	OCC_TEST
Data Collection Name	OCC_TEST

Errors Details

Type	Rule code	Error code	Error Description	Variables	Example
E	INSERT_FAIL	INSERT_FAIL	5 rows of the file : Example1.xls were not inserted (7081/8417)		
E	BR03A	ER14B	The result LOD must be less than the LOQ	resLOD$<=$resLOQ	1$<=$.8
E	BR03A	ES28B	Sample year cannot be greater than the analysis year	sampY$<=$analysisY	2011$<=$2010
					$RF-XXXX-XXX-

A priority from the start: Hierarchical rules

All data domains

Standard Sample Description ver. 1

General Validation Rules

Contaminants

Specific requirements for contaminants

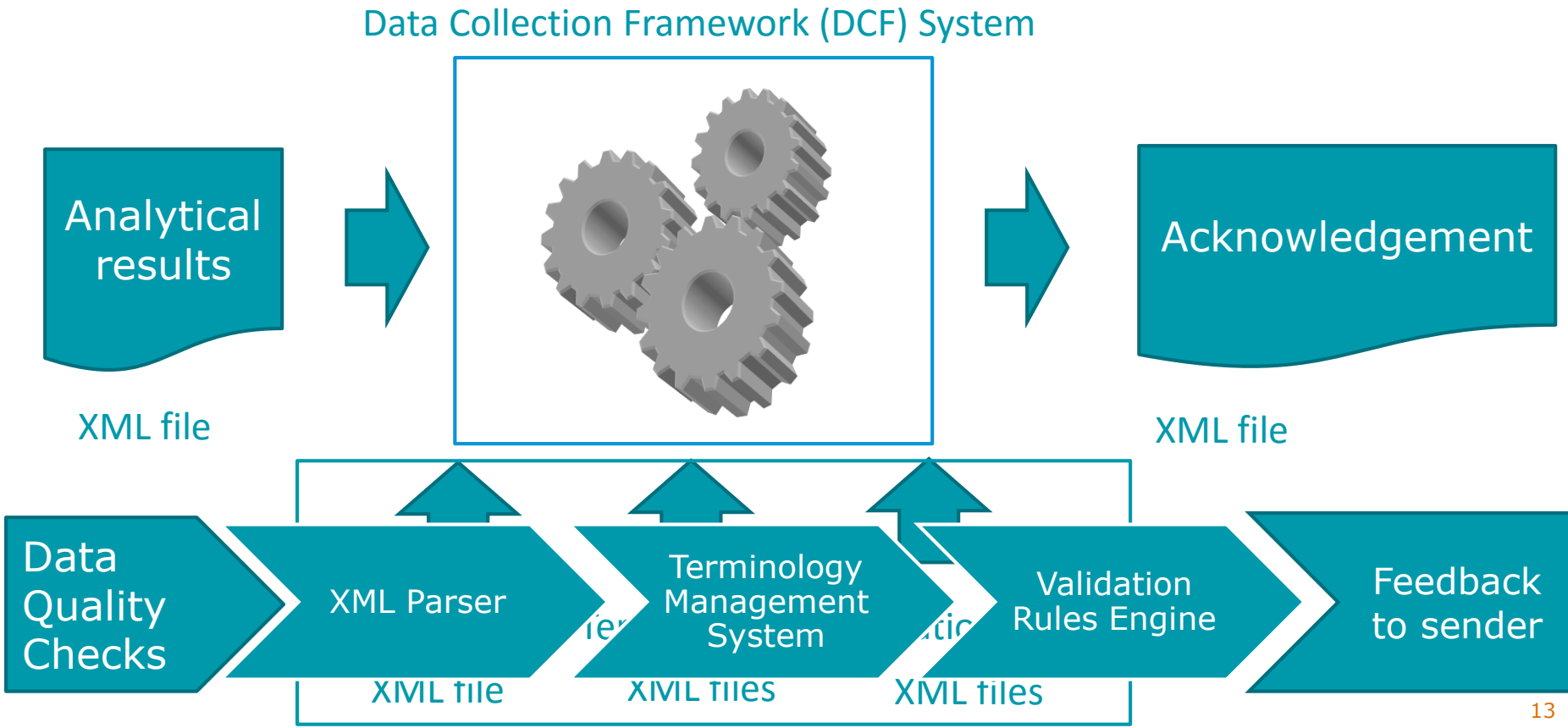
Specific validation rules for contaminants

Pesticide residues

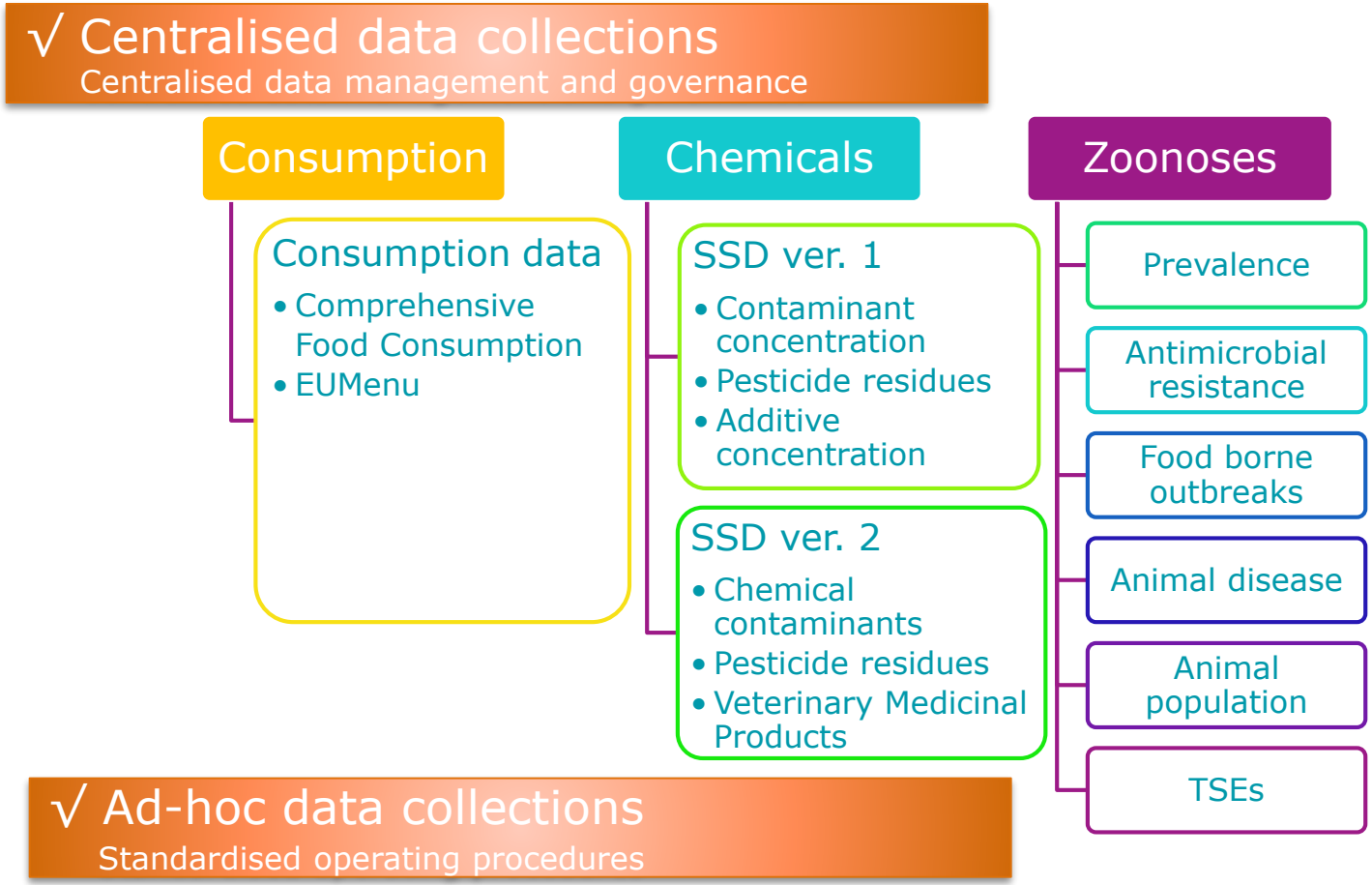
Guidance for pesticide residues

Specific validation rules for pesticide residues

A priority from the start: data validation

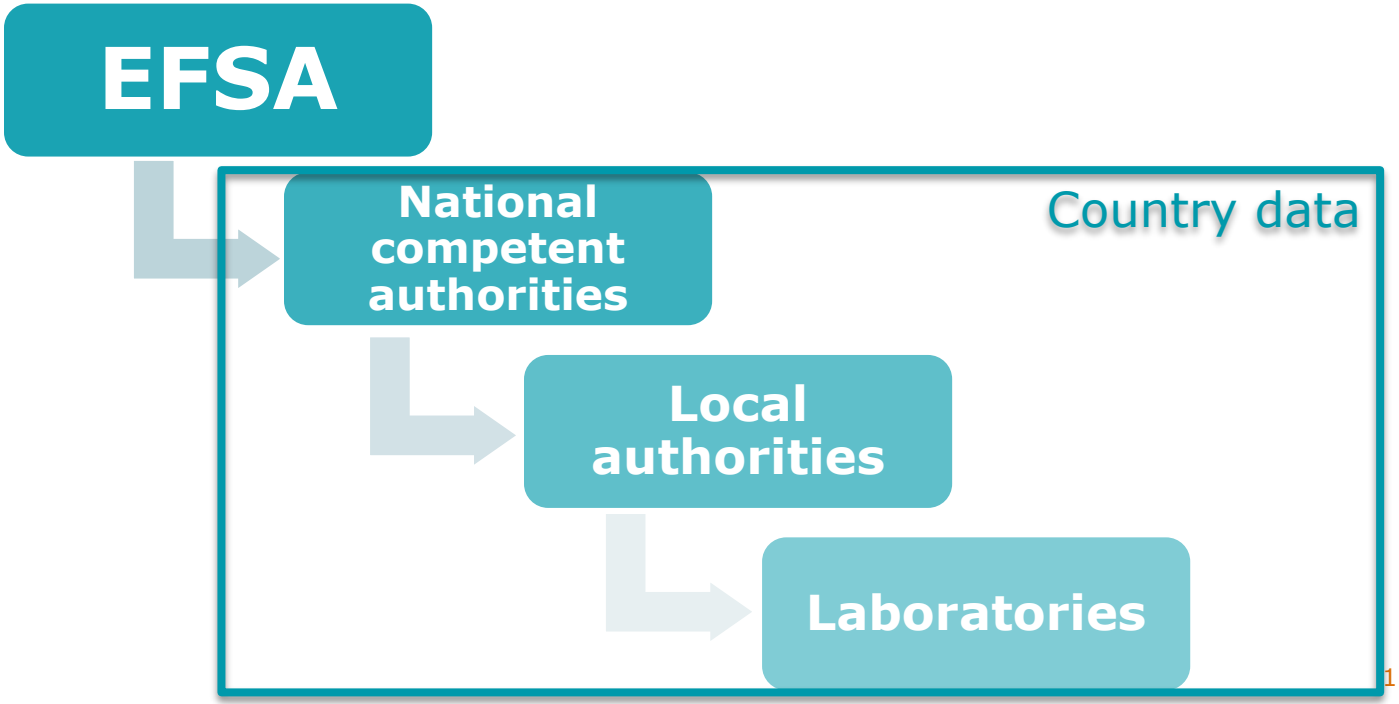


A priority from the start: Process standardisation...



A priority from the beginning…not only for EFSA

Standardisation of data reporting is across different domains and the entire data collection chain

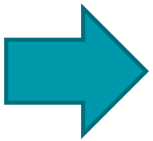


Summary

- Background
- Engagement from the start
- Quality of data
- Quantification of data quality
- Visualising data quality
- Governance and collaboration with Member States
- Conclusions

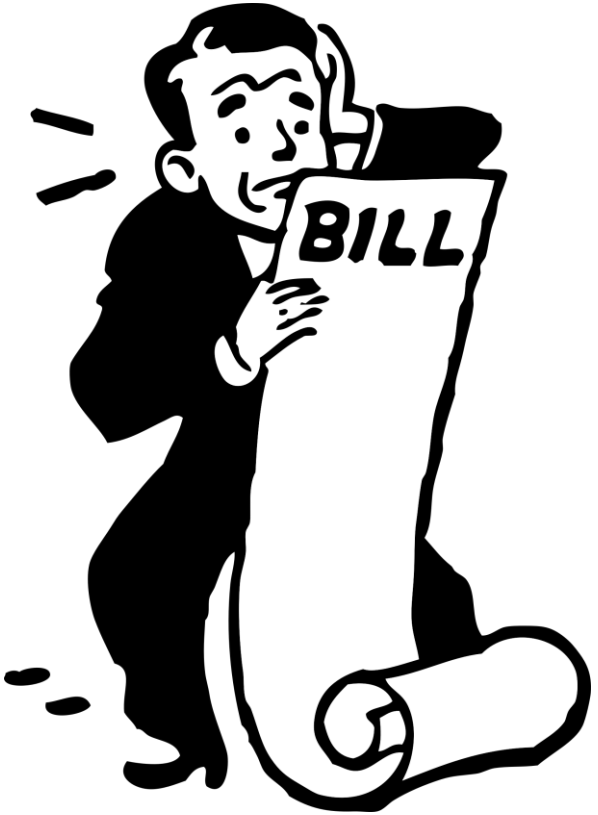
Quality of data: More data quality, please!

- Workload intensive
  - All data reporting levels are impacted for:
    - Resources
    - Costs
- Data users want more data quality



Can we actually use the data?

Quality of data



Quality of data: A definition

Data quality is the extent to which data are **fit for their intended use**

(in line with the general definition of quality as set in the standard ISO 9000: "degree to which a set of inherent characteristics of an object fulfils requirements")

Going to the next level: Data Quality Management Framework

Going to the next level: Data Quality Virtuous Cycle

Data Quality Management Framework

1. **DEFINE** use cases and requirements for data quality (**Data Quality Objectives**)
2. **MEASURE** quality of data by **Key Performance Indicators**

3. **ANALYSE** quality assessment **outcomes**
4. **IMPROVE** by taking corrective **actions**

This also matches the PDCA (Plan, Do, Check, Act) cycle of more general quality assurance process (ISO9001, 2016).

Summary

- Background
- Engagement from the start
- Quality of data
- Quantification of data quality
- Visualising data quality
- Governance and collaboration with Member States
- Conclusions

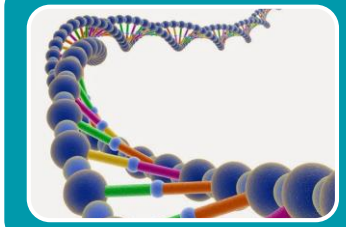
Quantification of data quality: DQ dimensions

- Data quality dimensions: Classification system of positive features often desired in the data, used to help specification of data quality objectives and aggregation, reporting and comparison of data quality analysis
  - Classification system: DQ dimensions classify features of data as food classes classify food (e.g. cereal and cereal products, vegetables, dairy products)
  - Positive: standardise to look positive features (mainly not to get confused)
  - Often: They do not need to be all present at all time

Quantification of data quality: DQ dimensions

Validity

- Are data elements consistent to their format, type and range?
- Are constraints respected?



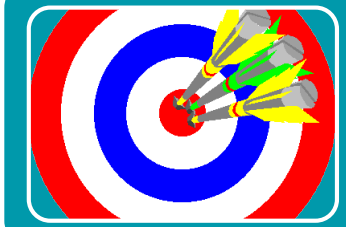
Uniqueness

- Are the records present only once in the database?
- Are database unique identifier available?

Timeliness

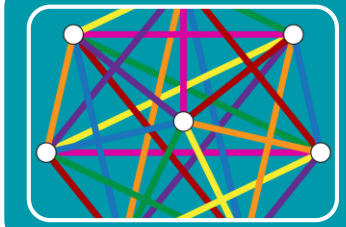
- Are data available when needed?
- Are data up to date for their uses?

Quantification of data quality: DQ dimensions



Accuracy

- Are data elements representing correctly the real world from which are extracted?
- Are the data plausible?



Completeness

- Is information reported in the data elements comprehensive?
- Are valuable data elements missing?

Consistency

- Are different data elements providing non-conflicting details for a specific piece of information?

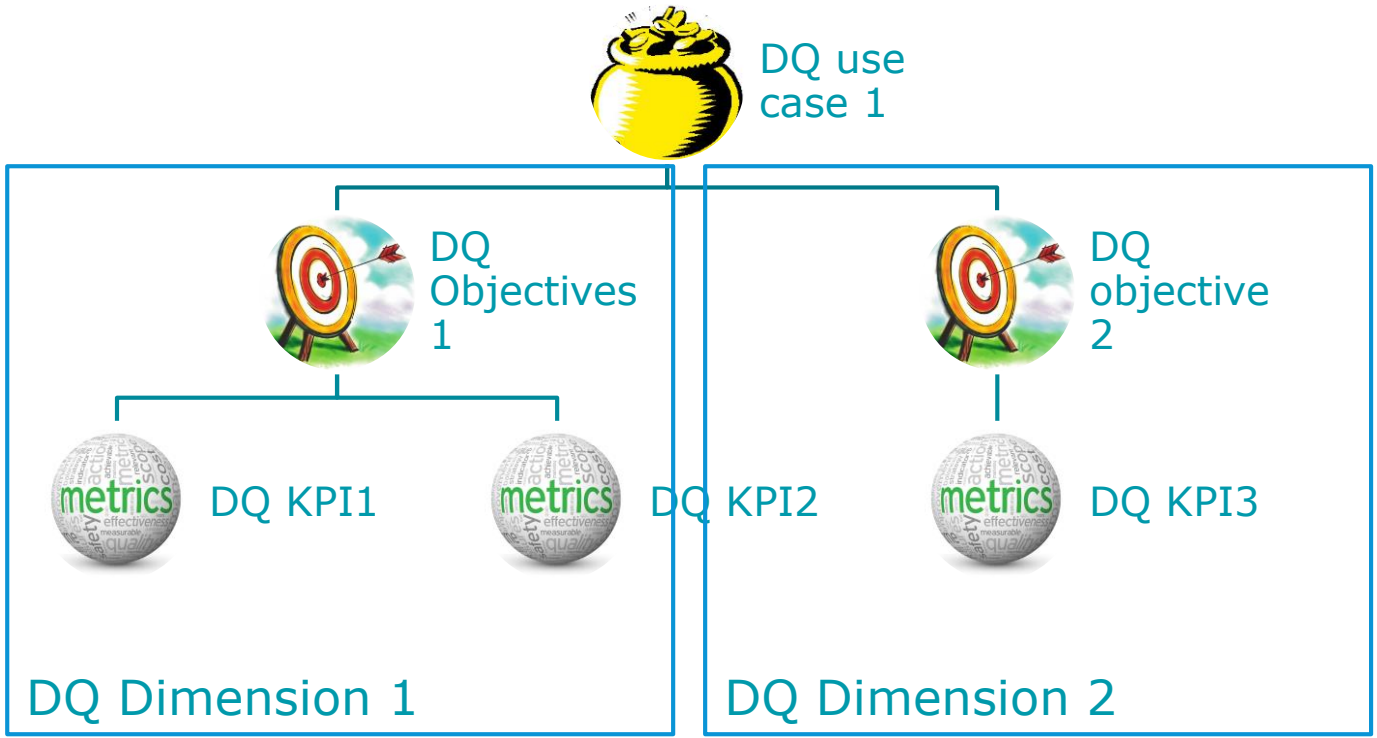
Quantification of data quality: DQ use cases and objectives

- **Data use cases**: Define the main uses for which data are collected and managed. For example:
  - Risk assessment, exposure assessment
  - Risk management
- **Data quality objectives**: Requirements that data users expect in the data, in order to fulfil their uses

Quantification of data quality: DQ KPIs and thresholds

- Data quality Key Performance Indicators(KPIs):
  - Indicators measuring the level of fulfilment of a data quality objective
- Data quality thresholds:
  - levels that the DQ KPIs must achieve in order to consider the DQ objective fulfilled.

Quantification of data quality: general overview



Quantification of data quality: general principles

1. Keep it simple
2. Focus on incoming data: Data Quality KPIs implemented and calculated on incoming data from as submitted by data providers. Data quality at entrance.
3. Only most relevant dimensions
4. Minimise KPIs: Potentially one for objective
5. Avoid complex formulas for KPIs: Use only proportions
6. Agree data quality objectives and KPIs with data providers

An example: chemical contaminants

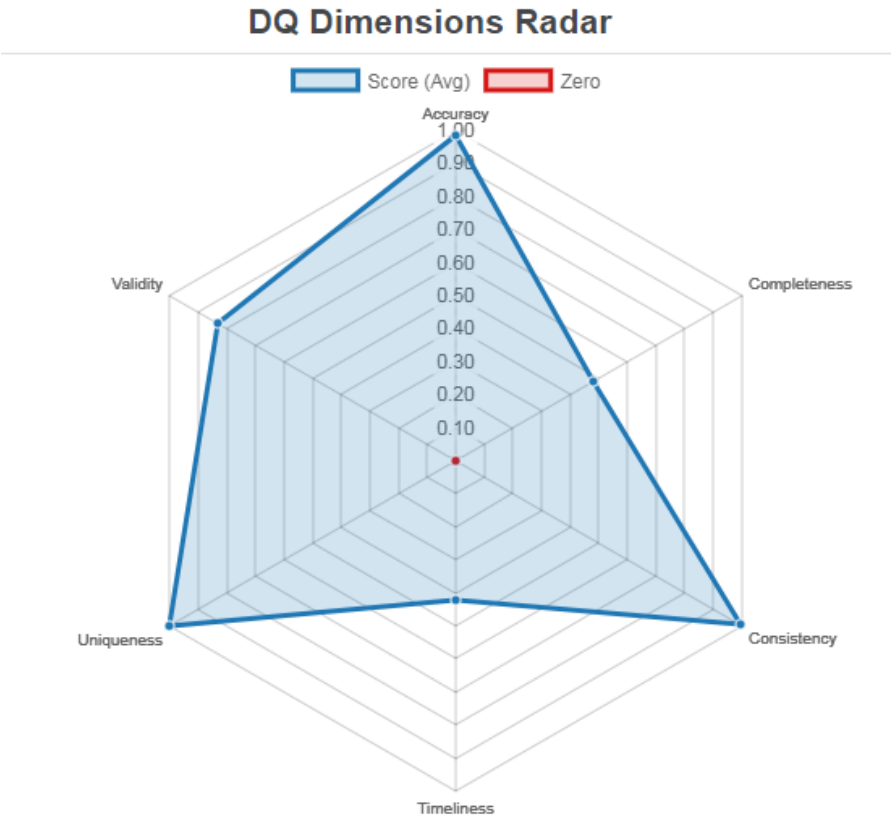
Quantification of data quality: define data quality objectives/KPIs

DQ Use case	DQ Objectives	DQ KPI
Risk assessment/exposure assessment	**DQO_CHEM_01**: Timely availability of the data for analysis to risk assessors and risk managers	**KPI_CHEM_01**: Proportion of data records in "SUBMITTED" status by data collection deadline
		KPI_CHEM_02: Proportion of data records confirmed by data providers within one month (calendar) from data collection deadline
	DQO_CHEM_03: No duplication of records	**KPI_CHEM_03**: Proportion of data records not duplicated in a data collection
	DQO_CHEM_06: No mistakes for relevant numerical values (e.g. VAL, LOD or LOQ)	**KPI_CHEM_04**: Proportion of records containing the correct numerical value for the analytical result (i.e. resVal)
		KPI_CHEM_05: Proportion of records containing the correct limit of detection for the analytical result (i.e. resLOD)
		KPI_CHEM_06: Proportion of records containing the correct limit of quantification for the analytical result (i.e. resLOQ)

Summary

- Background
- Engagement from the start
- Quality of data
- Quantification of data quality
- Visualising data quality
- Governance and collaboration with Member States
- Conclusions

Visualising data quality: contaminants DQ dimensions

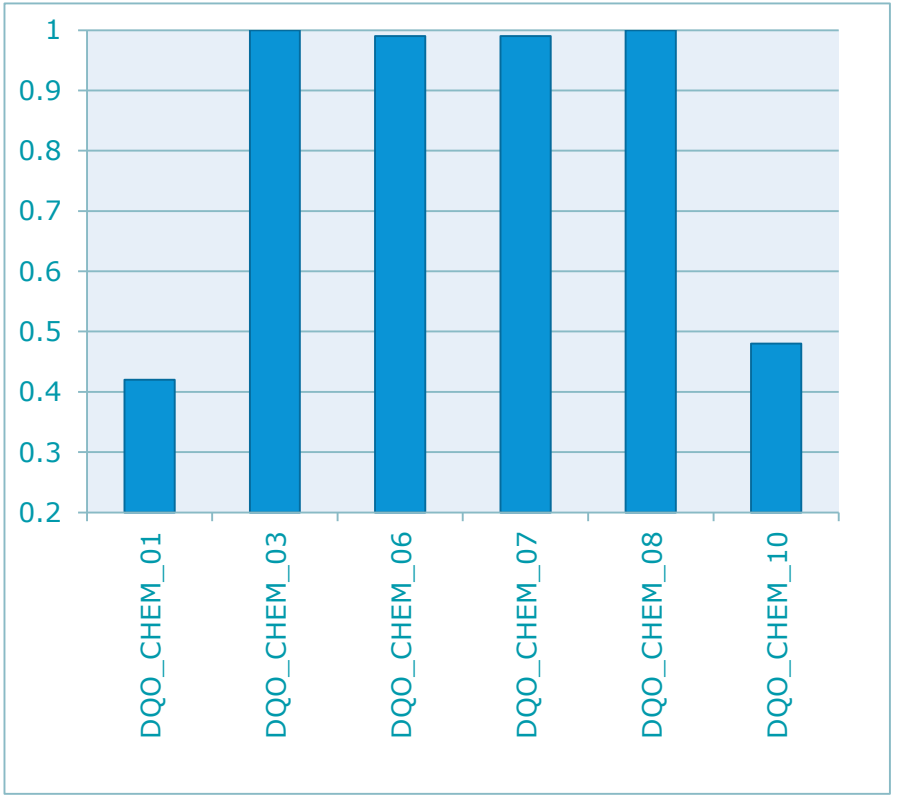


Data Quality Dimensions

Data Quality Dimensions	Overall Score
Accuracy	0.995
Completeness	0.371
Consistency	0.999
Timeliness	0.667
Uniqueness	1.000
Validity	1.000

Visualising data quality: Contaminants DQ Objectives

Metrics	
DQO_CHEM_01	0.42
DQO_CHEM_03	1.00
DQO_CHEM_06	0.99
DQO_CHEM_07	0.99
DQO_CHEM_08	1.00
DQO_CHEM_10	0.48



Visualising data quality: Contaminants DQ KPIs

DATA QUALITY KPIs in Chemical Contaminants DC

Metrics	
Timely Submission (KPI_CHEM_01)	0.52
Timely Confirmation (KPI_CHEM_02)	0.29
Uniqueness (KPI_CHEM_03)	1.00
ResVal Accuracy (KPI_CHEM_04)	0.98
Res LOD Accuracy (KPI_CHEM_05)	0.99
ResLOQ Accuracy (KPI_CHEM_06)	0.99
Consistent Food Description (KPI_CHEM_07)	0.99
Consistent Substance Description (KPI_CHEM_08)	1.00
No Generic Terms (KPI_CHEM_09)	0.46
SSD Submission (KPI_CHEM_10)	0.80

Visualising data quality: Contaminants Drill down completeness

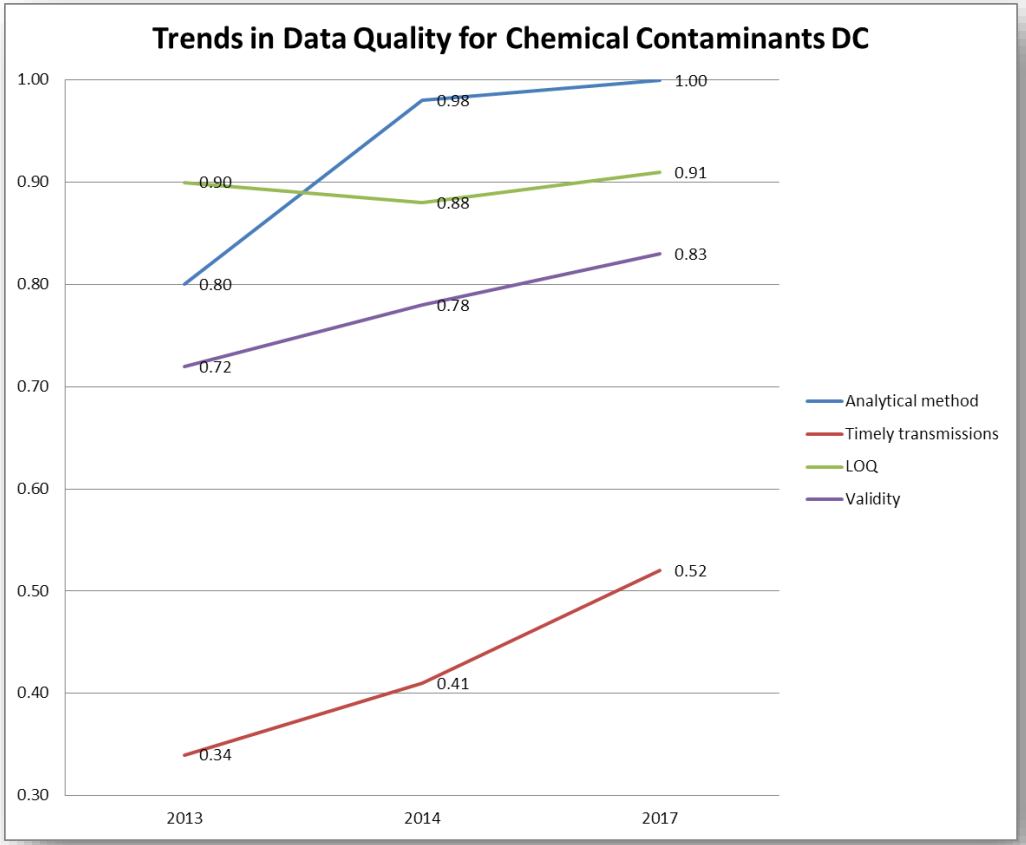
Completeness: generic terms

Terms	Score
Foodex	1.00
Analytical Method	1.00
Country of Sampling	1.00
Country of Origin	0.84
Sampling Method	0.93
Sampling Point	0.94
Program Type	0.99
Sampling Strategy	0.91
Reported LOQ	0.91
Product Treatment	0.67

Unknown
Not Reported
Not Available
...

Not Reported
Not Available
Unknown

Visualising data quality: contaminants trends



Trend in Data Quality since 2013

Validity ↑

Timeliness (submission) ↑

Analytical method ↑

LOQ reported →

Visualising data quality: DQ actions

- **Timeliness**

Streamline communication - clear rules must be defined and disseminated by EFSA on how to send data, by when, and how to perform data confirmation

- **Completeness**

Re-discuss the inclusion of generic terms in the catalogues related to the data elements highlighted:

Mandatory + generic term (e.g. “Unspecified”) = Optional

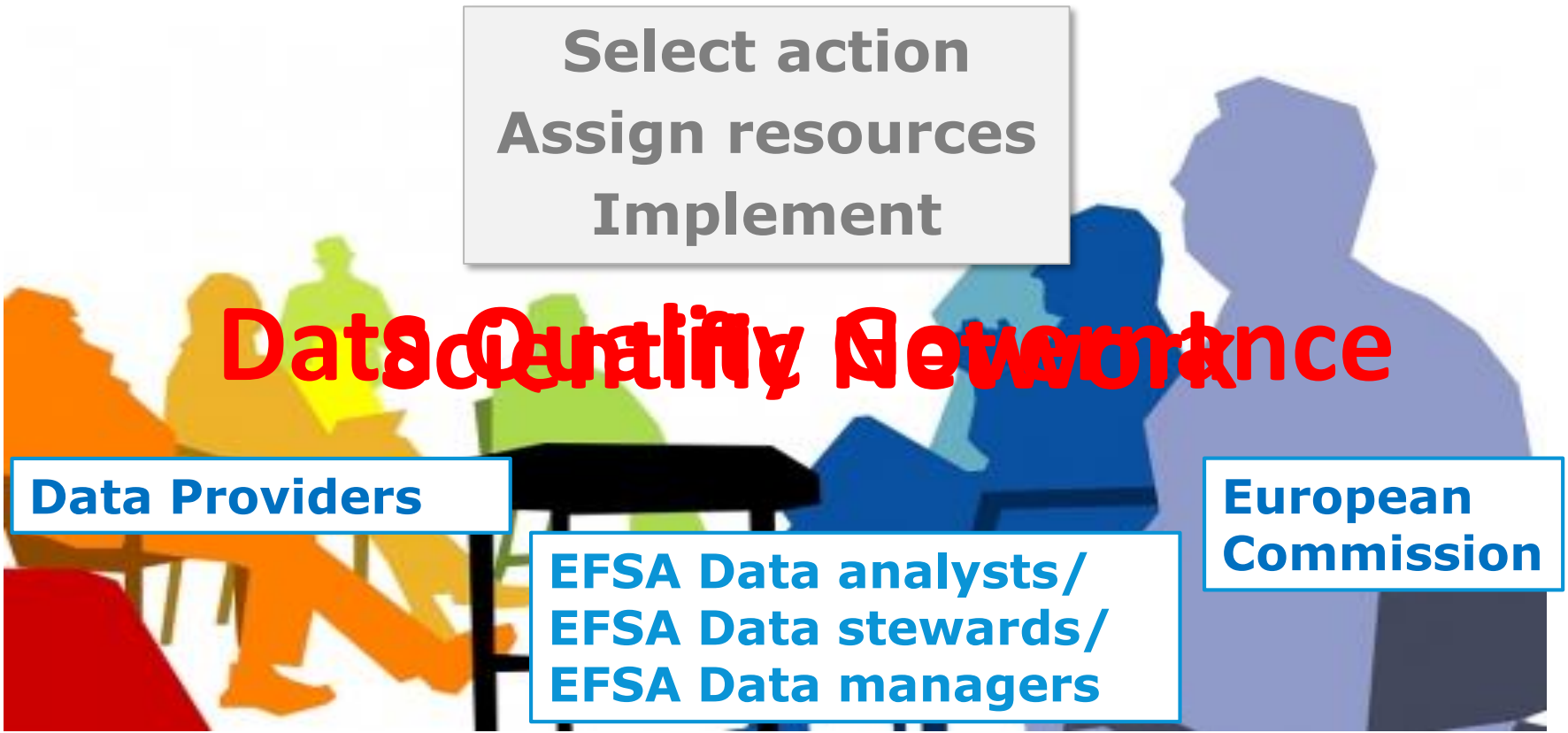
- **Validity**

Promote automation of transmissions to reduce non standard format (20%)

Summary

- Background
- Engagement from the start
- Quality of data
- Quantification of data quality
- Visualising data quality
- Governance and collaboration with Member States
- Benefits, limitations and lessons learnt

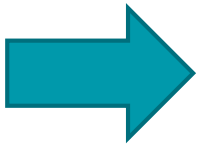
Governance and collaboration with MSs



Governance and collaboration with MSs

Framework Partnership Agreement on Data Quality

- Support definition of DQ Objectives and KPIs
- Agree actions for data quality improvement
- Coordinate data stewardship activities to improve data quality, and cross domain issues
- Decide priorities to invest available resources for data stewardship and data management automation



Pilot of a Framework Partnership Agreement (FPA) on Data Quality

Governance and collaboration with Member States

Framework Partnership Agreement on Data Quality: Objectives

Governance and collaboration with Member States

Framework Partnership Agreement on Data Quality

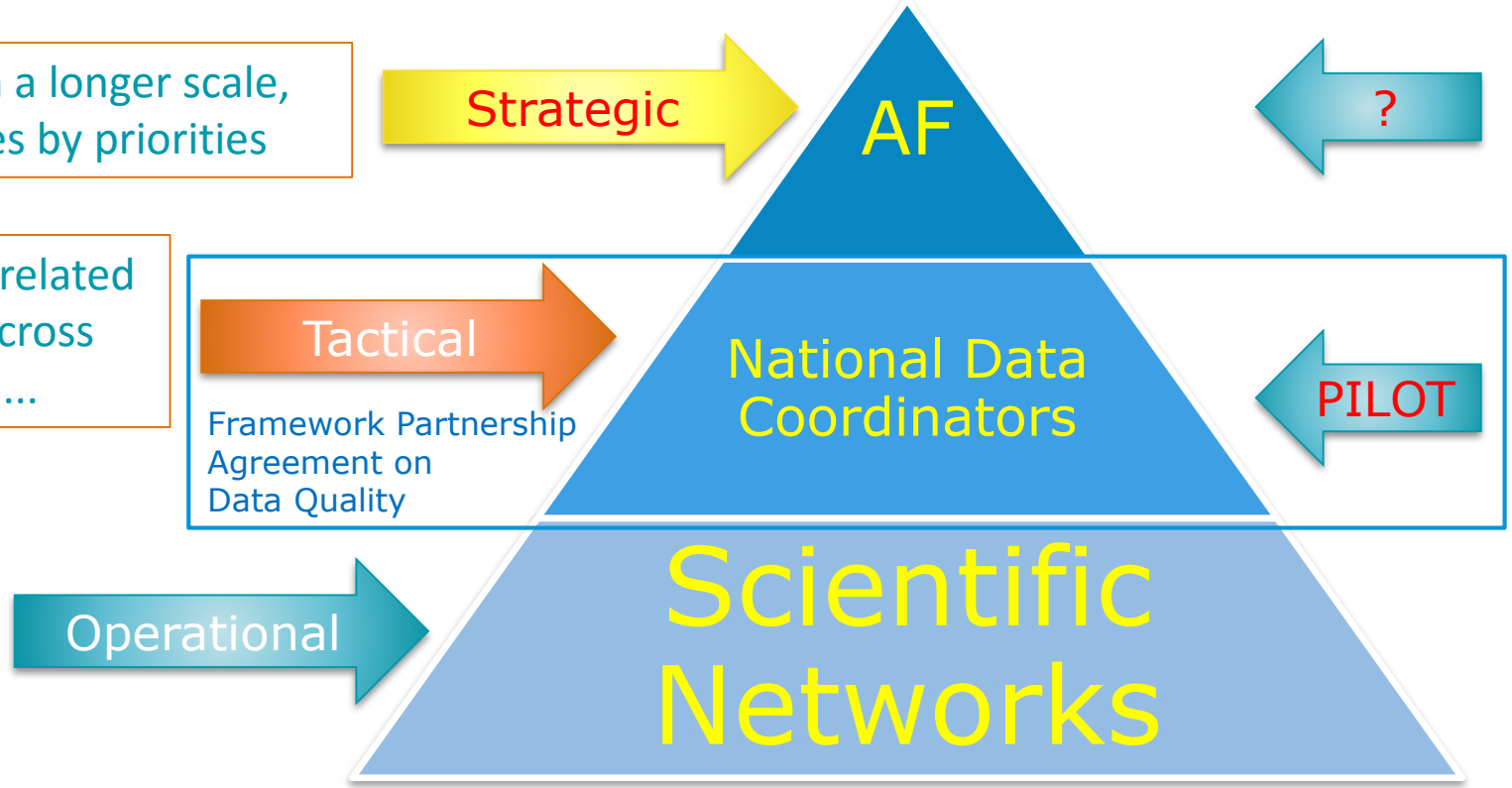
- Pilot
- Five countries (Cyprus, Denmark, France, Germany, Slovakia) selected by geographic distribution and size
- Essential to negotiate data quality objectives and KPIs
- Additional network members were involved in the discussions

Goverance and collaboration with MSs: Data governance Pyramid

Specify needs on a longer scale, allocate resources by priorities

Broader aspects related to Data Models, cross domain aspects, ...

Define DQO, select KPIs, Propose actions



Summary

- Background
- Engagement from the start
- Quality of data
- Quantification of data quality
- Visualising data quality
- Governance and collaboration with Member States
- Conclusions

Benefits

- The effort for improving data quality is connected to actual value (DQ use cases) for the data users
- The need for certain actions of data quality improvement are evident through the process of DQ Objectives ad KPIs (e.g. adding a new business rule)
- Make evident the cost of data quality, define limits
- Provide a governance of the process o where actions and responsibilities are shared with data providers (e.g. Member States)

Constraints

- DQ Framework applied only to incoming data and not to the entire EFSA Scientific Data Warehouse
- DQ Framework applied only to data passing the validation phase. The framework should be extended to the entire set of applied validation rules
- Only National Competent Authorities involved so far: particularly from FPA pilot. Reactions from other data providers must be investigated

Lessons learnt

- Engagement of data providers is essential
- Start small (few dimensions, objectives and KPIs)
- Keep It Simple approach for KPIs
- Low score in KPIs is an effect and not a cause: always explain it to Data Providers
- Be ready to clean the house on recipient side:
  - Clarify Service Level Agreements on receiver side

Acknowledgements

- Valentina Bocca: Data Quality/Data validation
- Alessandro Carletti: Data Quality/Data Quality Framework

Any questions?

References

References: Standard Sample Description (SSD)

- SSD ver 1.0
  - European Food Safety Authority; Standard sample description for food and feed. EFSA Journal 2010;8(1):1457 [54 pp.].doi:10.2903/j.efsa.2010.1457.
- SSD ver 2.0
  - EFSA (European Food Safety Authority), 2013. Standard Sample Description ver. 2.0. EFSA Journal 2013;11(10):3424, 114 pp., doi:10.2903/j.efsa.2013.3424

References: Contaminants Requirements

- For SSD1 data format:
  - EFSA (European Food Safety Authority),2017. Specificreporting requirements for contaminants and foodadditives occurrence data submission.EFSA supportingpublication 2017:EN-1262. 27 pp. doi:10.2903/sp.efsa.2017.EN-126
- For SSD2 data format:
  - EFSA (European Food Safety Authority),2017. Specific reporting requirements for contaminants and food additives occurrence data submission in SSD2. EFSA supporting publication 2017:EN-1261. 43pp. doi:10.2903/sp.efsa.2017.EN-1261

References: Pesticide residues requirements

- Yearly updated guidance, publication 2017
- For SSD1 data format:
- EFSA (European Food Safety Authority), Brancato A, Brocca D, Erdos Z, Ferreira L,Greco L, Jarrah S, Leuschner R, Lythgo C, Medina P, Miron I, Nougadere A, Pedersen R, Reich H, Santos M,Stanek A, Tarazona J, Theobald A and Villamar-Bouza L, 2017. Guidance for reporting data on pesticideresidues in food and feed according to Regulation (EC) No 396/2005 (2016 data collection). EFSA Journal2017;15(5):4792, 48 pp.https://doi.org/10.2903/j.efsa.2017.4792

References: Veterinary Med. Prod. reporting requirements

- Quick start reporting guide ver 2
  - https://zenodo.org/record/1204115#.WvBfby5ubcv

Contaminants KPIs (1)

Data Quality Objective	Description
DQO_CHEM_01	Timely availability of the data for analysis to risk assessors and risk managers. Late or last minute data transmissions delays availability of data to support risk assessment and management processes. In addition it increases the risk of not identifying and fixing possible data quality issues.
DQO_CHEM_02	Timely availability of data updates. These requests are part of broader workflows for the use of data. The time for answering influences the entire process
DQO_CHEM_03	No duplication of records. Finding, resolving and reducing the incidence of duplicated records in delivered datasets by data providers. Duplicated records impede the suitability of the data for immediate use and, if not identified, put at risk the value of the analysis.
DQO_CHEM_05	Completeness of the dataset with respect to the planned and performed analyses and inclusion of all results. Omitting results from the original plan reduces the representativeness of the data and not reporting some results seriously biases the statistics on the occurrence of the hazards and consequently the exposure estimate.
DQO_CHEM_06	No mistakes for relevant numerical values (e.g. VAL, LOD or LOQ) or in the associated unit of measurement. Finding, resolving and reducing the incidence in records of numerical errors in delivered datasets by data providers. Records with numerical errors impede the suitability of the data for immediate use and, if not identified, put at risk the value of the analysis.

Contaminants KPIs (2)

Data Quality Objective	Description
DQO_CHEM_07	Detailed and consistent identification of the analysed Matrix (e.g. food, feed) coded according to the relevant food classification system. A detailed identification of the matrix allows a better granularity of the data analysis thus improving the assessment; A detailed and correct use of the matrix catalogue expedite the direct use of the collected data for analysis and reduces the time needed for manual data cleansing preparing the data for risk assessment.
DQO_CHEM_08	Precise and consistent identification of the Parameter (analyte) not using generic browsing terms. The Parameter catalogue is a multi-level catalogue and the aggregated terms are useful for navigating the hierarchy, but only the use of detailed terms allows a precise data analysis
DQO_CHEM_09	Pertinence and correctness of the expression of the result for the combinations parameter-matrix. The matrix reported must be relevant for the parameter analysed and must be expressed in the correct expression of the result.
DQO_CHEM_10	No incomplete data for mandatory and recommended elements. Different catalogues also include generic descriptors (e.g. other, unspecified, not in list), but the use of such descriptors seriously reduces the usability of data.
DQO_CHEM_11	No data provided outside agreed standard format (SSD1 o SSD2). Chemical contaminants data are often provided in non-standard format that can cause various transcription and processing issues.