

tds ► exposure

WP6/WP10 Workshop Data Quality Framework

**WP6/WP10 Workshop
Zurich, 13th February 2015**



Agenda

- ▶ Data Quality Framework (DQF) concept
 - Constraints (rules) definition
 - Contracts
- ▶ Data input validation & data quality analysis
- ▶ Measurement and calculation of data quality
- ▶ Data quality visualisation
- ▶ Comparison to Food Composition DQF
- ▶ Your feedback / discussion
- ▶ Survey



Data Quality Framework (DQF) concept

- ▶ Managing and measuring data quality (DQ)
- ▶ Constraints as DQ indicators
- ▶ Unique definition and validation
- ▶ Flexible and highly configurable DQF
 - Importance weights for constraints
 - Grouping of constraints
 - DQ contracts for different users
 - Various visualisations



DQF concept - Constraint definition

- ▶ Constraints defined in a single place in the system (by developer)
- ▶ Constraint groups can be defined (by developer)
 - Constraints can belong to one or more groups
 - Allow organising constraints into logical units (Constraint types, accuracy, completeness, etc.)
 - Allow to validate/analyse only certain groups



DQF concept - Constraint definition cont.

- ▶ No possibility to add new constraints or constraint groups for a user of the application
- ▶ User can define weights and threshold
 - 0: Not important, 100: Most important
 - Threshold between warning and error
 - Default weight
- ▶ Weights can be overwritten (if configured)

DQF concept - Contracts

- ▶ A contract defines the weight (importance) for each of the constraints
- ▶ A contract is a simple file (editable in GUI)
- ▶ Default contract exists as fallback
- ▶ Users can define their (multiple) own contracts
- ▶ Contracts can easily be imported and shared between users



DQF concept - Example

- ▶ Sample: Kartoffelsuppe (engl. Potato soup)
- ▶ AnalysisValue: Total copper, 0.845 mg/kg
- ▶ Two constraints:
 - Sample.englishName.NotEmpty=80, active
(very important)
 - AnalysisValue.verifyingBy.NotEmpty=20, active
(less important)
- ▶ Other constraints: Min/Max, Range, Email, URL, Pattern, etc. and custom constraints ('Was the subsample bought in the store as planned?')

Data input validation

fc TDS pool detail, pool id: 6 / import source id: Import source

File

Save Save & close Close

TDS Pool

Id 6

Pool name Knoblauch

English name Garlic

Scientific name No scientific name

Load picture...

Pool Info Pool Names LanguaL FoodEx2 Markers Sampling Remarks Categories Samples Files Pooling

FoodEx2 Exposure Garlic and similar (p) (A00GY)

FoodEx2 Facet

Food Ex2 Facet

Add...

Remove

FoodEx2 comment

Comment

Data quality evaluation

Well done, currently there are no constraint violations!
Current object quality is 100.0%

Data input validation

fc TDS pool detail, pool id: 6 / import source id: Import source

File

Save Save & close Close

TDS Pool

Id 6

Pool name

English name Garlic

Scientific name No scientific name

Load picture...

Pool Info Pool Names Language FoodEx2 Markers Sampling Remarks Categories Samples Files Pooling

FoodEx2 Exposure Garlic and similar (p) (A00GY)

FoodEx2 Facet

Food Ex2 Facet

Add...

Remove

FoodEx2 comment

Comment

Data quality evaluation

The pool name cannot be empty.
The pool name must have at least 2 characters.
Current object quality is 59.523808%

Data input validation

FC TDS Pool value detail

File

Save Save & close Close

Pool name Knoblauch Value id 6

Substance name Salmonella Worb

Value 0.06 Value type (MN) mean

Unit (mg) milligram Matrix unit (W) per 100g edible portion

Value Lab analysis Markers Remarks Files

Measurement limits

LOD 0.0002 LOQ 0.0003

LOR 0.0001

Statistics

Mean 0.06 Median 1

Minimum 0 Standard deviation 0.01

Maximum 2 n 12

Uncertainty 0.1 Maximum uncertainty 0.2

Date

Verifying date 2014-07-30 Verifying by david

Data quality evaluation

Well done, currently there are no constraint violations!

Current object quality is 100.0%

Data input validation

FC TDS Pool value detail

File

Save Save & close Close

Pool name Knoblauch Value id 6

Substance name Salmonella Worb

Value 0.06 Value type (MN) mean

Unit (mg) milligram Matrix unit (W) per 100g edible portion

Value Lab analysis Markers Remarks Files

Measurement limits

LOD 0.0002 LOQ 0.0003

LOR 0.0001

Statistics

Mean 0.06 Median 1

Minimum 0 Standard deviation 0.01

Maximum 2 n 12

Uncertainty 0.1 Maximum uncertainty 0.2

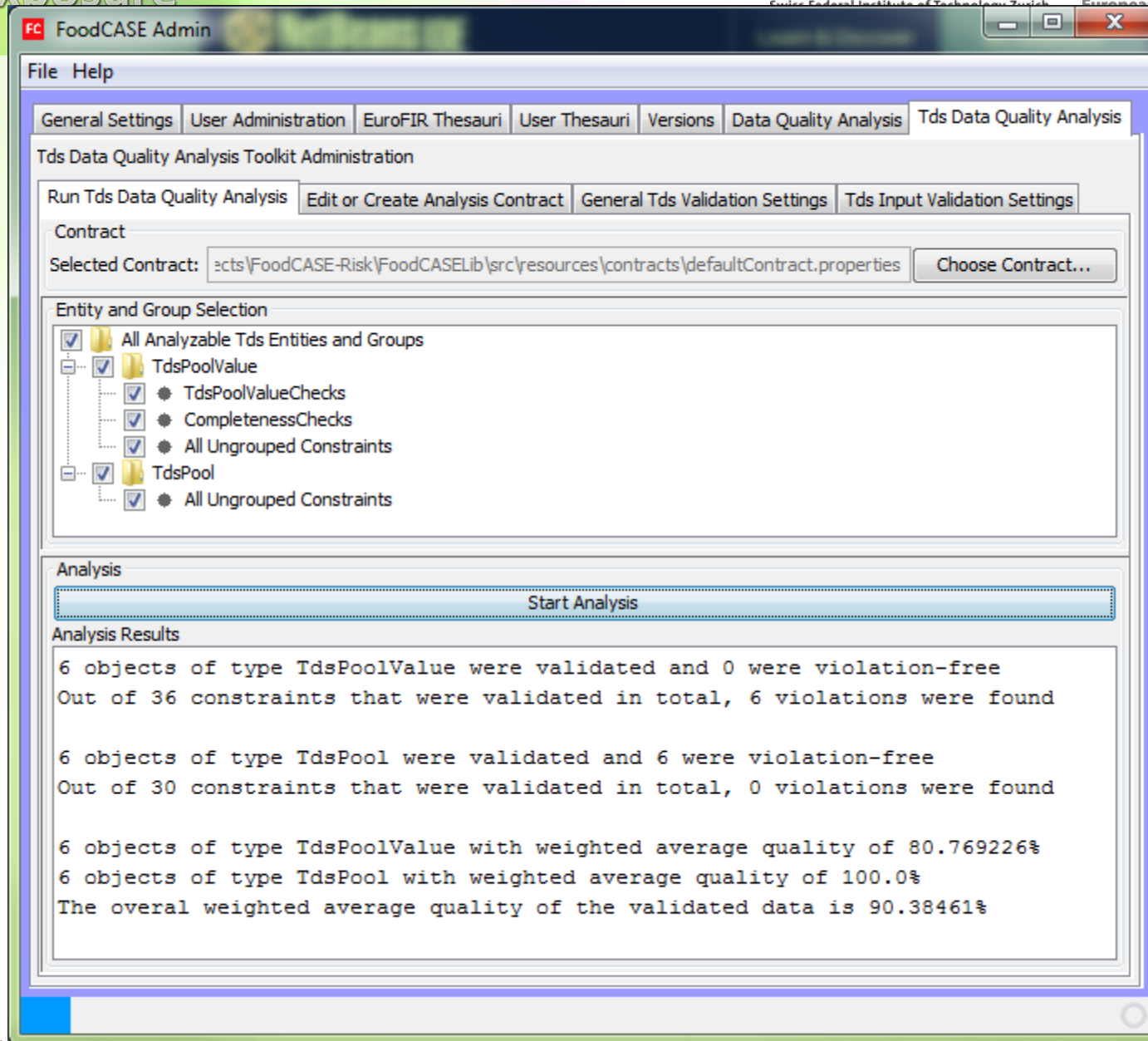
Date

Verifying date 2014-07-30 Verifying by

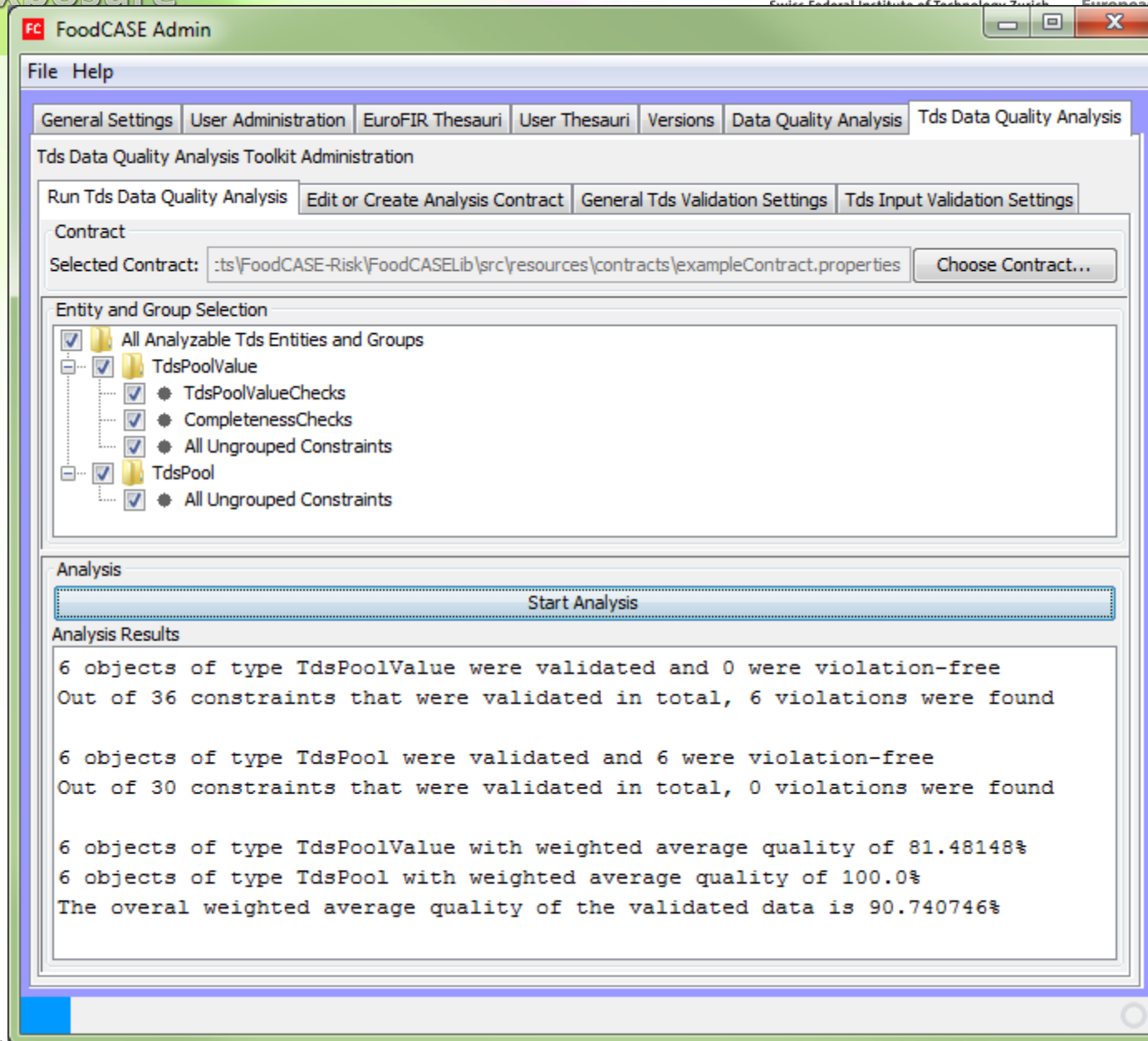
Data quality evaluation

'Verifying by' field should not be empty.
Current object quality is 92.30769%

Data quality analysis



Data quality analysis



Data quality analysis

FoodCASE Admin

File Help

General Settings User Administration EuroFIR Thesauri User Thesauri Versions Data Quality Analysis Tds Data Quality Analysis

Tds Data Quality Analysis Toolkit Administration

Run Tds Data Quality Analysis Edit or Create Analysis Contract General Tds Validation Settings Tds Input Validation Settings

Contract

Selected Contract: s:\FoodCASE-Risk\FoodCASELib\src\resources\contracts\defaultContract.properties Choose Contract...

Edit Contract

Entity Name	Entity Field (optio...	Constraint Name	Weight	Has Payload	Is Inactive
TdsPool	name	Size	90	☐	☐
TdsPool	englishName	NotEmpty	100	☐	☐
TdsPool		PoolFatConsistent	90	☐	☐
TdsPool		UniquePoolName	100	☐	☐
TdsPool	name	NotEmpty	100	☐	☐
TdsPoolValue		ValidMinMax	80	☐	☐
TdsPoolValue	minimum	Min	100	☐	☐
TdsPoolValue	value	Min	100	☐	☐
TdsPoolValue	verifyingBy	NotEmpty	30	☐	☐
TdsPoolValue	maximum	Min	100	☐	☐
TdsPoolValue		ValidRangeValue	80	☐	☐

Reload Contract New Contract... Save Changes

Data quality analysis

FoodCASE Admin

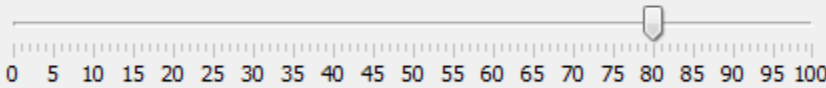
File Help

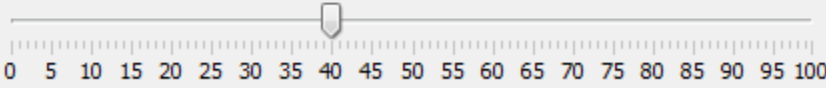
General Settings User Administration EuroFIR Thesauri User Thesauri Versions Data Quality Analysis Tds Data Quality Analysis

Tds Data Quality Analysis Toolkit Administration

Run Tds Data Quality Analysis Edit or Create Analysis Contract General Tds Validation Settings Tds Input Validation Settings

☒ Override constraint contract weight if its payload attribute is set

Default constraint weight: 

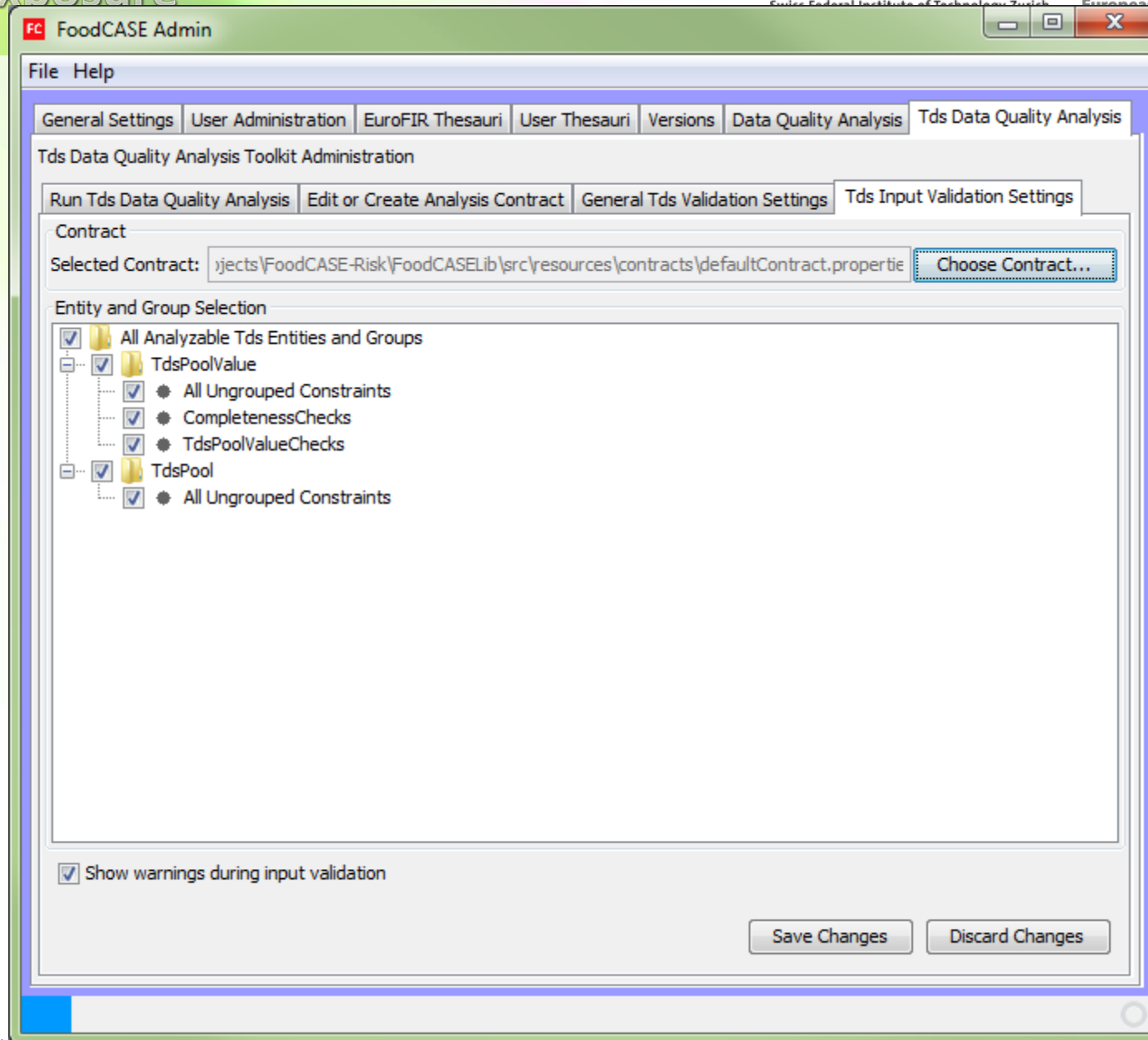
Warning to error threshold: 

Information about quality metric and scale

Here we will explain how quality is calculated, and describe the metric and scale that is being used for the calculations.

Save Changes Discard Changes

Data quality analysis



Data quality measurement / calculation

- ▶ DQ is calculated as the weighted arithmetic mean of all DQ indicators
- ▶ Other calculation schemes are possible but not implemented yet
- ▶ Users cannot influence the calculation directly but indirectly by defining the constraint weights



DQ calculation example

- ▶ 100 data entries, error threshold = 60
- ▶ Constraint A ('Sample English name not empty'), weight 80
- ▶ Constraint B ('Analysis value verified by notempty') , weight 20
- ▶ 5 entities violating constraint A (quality indicator A: 0.95)
- ▶ 40 entities violating constraint B (quality indicator B: 0.60)
- ▶ Overall entity quality indicator calculated as:

$$\frac{80 * 0.95 + 20 * 0.6}{80 + 20} = 0.88$$

weight A

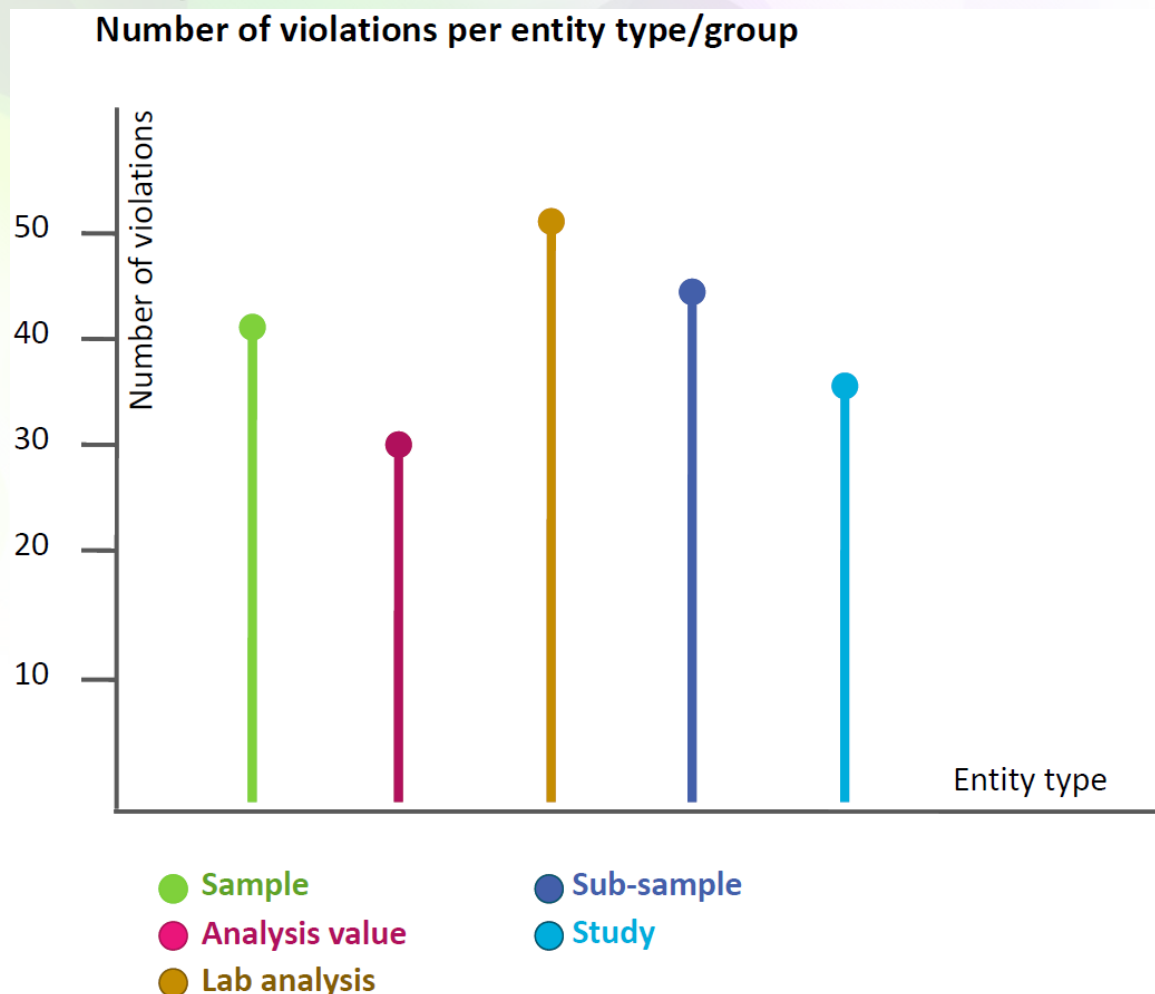
weight B

QI A

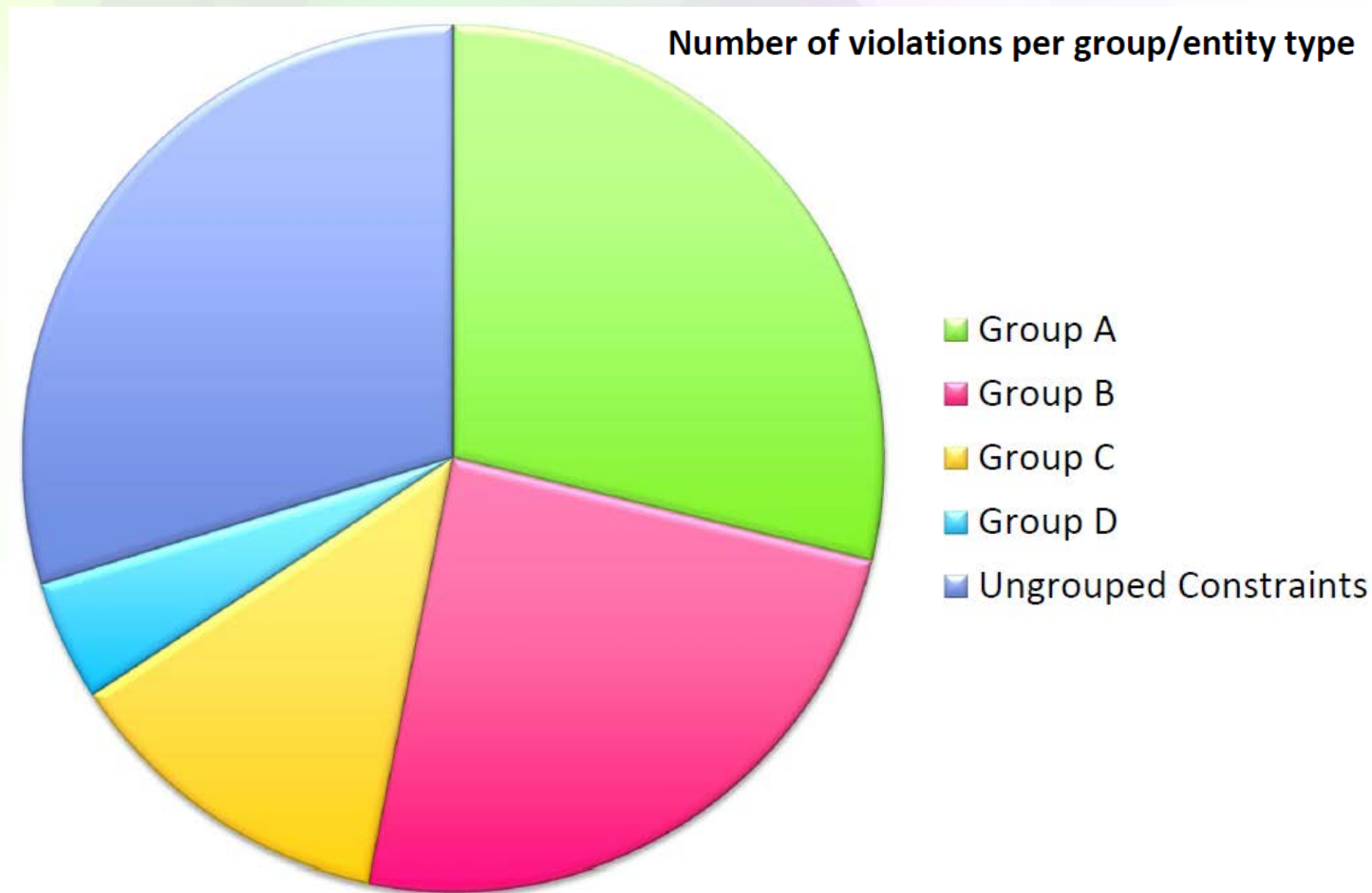
QI B



Data quality visualisation - Outlook



Data quality visualisation - Outlook



Data quality visualisation - Outlook

Number of constraint per weight range



Data quality visualisation - Outlook

25 violations of constraints with weight in the range of 80-100

- 10 violations of TdsPool.PoolFatConsistent (weight=80)
- 5 violations of TdsPool.englishName.NotEmpty (weight=80)
- 5 violations of TdsPoolValue.MinMaxValue (weight=90)
- 3 violations of TdsPoolValue.Value.NotEmpty (weight=100)
- 2 violations of TdsPoolValue.maxValue.Min (weight=100)

The maximum analysis value must be equal or larger than 0



Data quality visualisation - Outlook

Pool ID	Pool name	Study Name	Pool Value Id	Substance name
5	Brötchen	TDS 2011	5	Thiometon sulfoxide
7	Schokoladenkuchen	TDS 2011	4	Sulfosomidin

TDS Pool value detail

File Save Save & close Close

Pool name Brötchen Value id 5

Substance name Thiometon sulfoxide

Value 1 Value type (M) mean

Unit (mg) milligram Matrix unit (M) per 100g edible portion

Value Lab analysis Markers Remarks Files

Measurement limits

LOD 0.0002 LOQ 0.0003

LOR 0.0001

Statistics

Mean 0.05 Median 1

Minimum 0 Standard deviation 0.01

Maximum -1 n 12

Uncertainty 0.1 Maximum uncertainty 0.2

Date

Verifying date 2014-07-30 Verifying by david

Data quality evaluation

The maximum value must be greater than or equal to 0.
The minimum value can't be greater than the maximum value.
The pool value should be between minimum and maximum values.
Current object quality is 50.0%

Comparison to Food Composition DQF

- ▶ Input validation and DQ analysis
 - based on contracts (configurable individual importance weights per contract/user/user group)
 - user has influence on what is validated/measured
→ entities, groups
 - user has influence on the importance of constraints also for the input validation
 - Additional DQ feedback of importance (%)

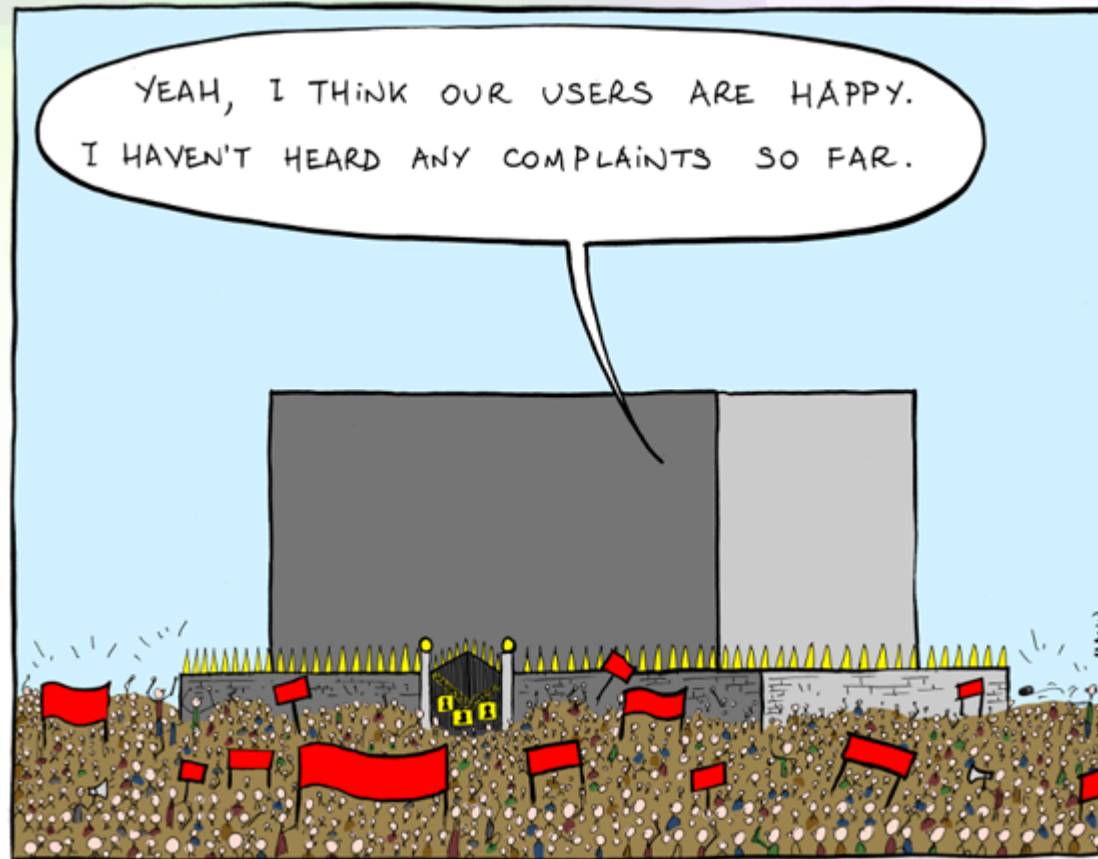


Comparison to Food Composition DQF

- Constraints
 - Constraints can be set as inactive
 - Configuration of default weight/importance
 - Configuration of warning error threshold
 - Warnings can be disabled
- Visualisation is open



Your feedback / discussion



Survey

- ▶ Please fill out all questions if possible
- ▶ Don't hesitate to ask if something is unclear
- ▶ If you have additional comments, ideas or requirements (such as a different DQ calculation) please let us know at the end of the questionnaire
- ▶ Time: 45 min
- ▶ Additional online survey for attributes planned



Thank you for your participation and attention

