

tds ► exposure



FoodCASE-Risk in the context of food data

David Weber (ETH Zurich)

TDS-Exposure Summer School, Helsinki, 9th of July 2015

David Weber

- ▶ MSc in Computer Science with focus on information systems and business informatics
- ▶ Several years of work experience as software engineer in the financial industry
- ▶ PhD student at ETH Zurich
- ▶ 3 years experience with food composition and TDS data management system
- ▶ Research interest on data quality

Outline

- Why to manage TDS data
- What data needs to be managed
- How data needs to be managed
- Architecture and food data entities
- Use cases in FoodCASE
- Data quality in FoodCASE
- [Concepts for data management]

Why to manage TDS data?

Why to Manage TDS Data

- ▶ Preserve data for long-term access
- ▶ Enhance data security
- ▶ Minimise risk of data loss
- ▶ Save time and resources in long run
- ▶ Enable other to access your data (unique way)
- ▶ Increase the visibility of your work
- ▶ Facilitate new discoveries

Why to Manage TDS Data

- Meet grant requirements
- See data as a product for which funding were used
- Open data means that publicly funded data must be publicly available (EU directive)

Take Home Message I

TDS data is the output of your
work and your asset.
Take care of it.

What Data Needs to be Managed?

How does a TDS work?



Food 1
Food 2
Food 3
...

plan

Collect food
according to
consumption
info

Food A
Food B
Food C
...

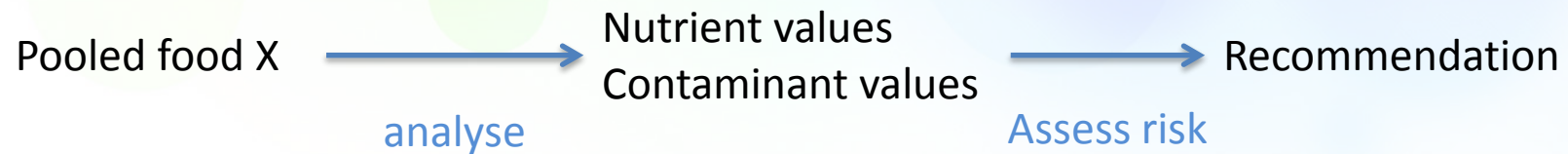
prepare as
consumed

Food A
Food B
Food C
...

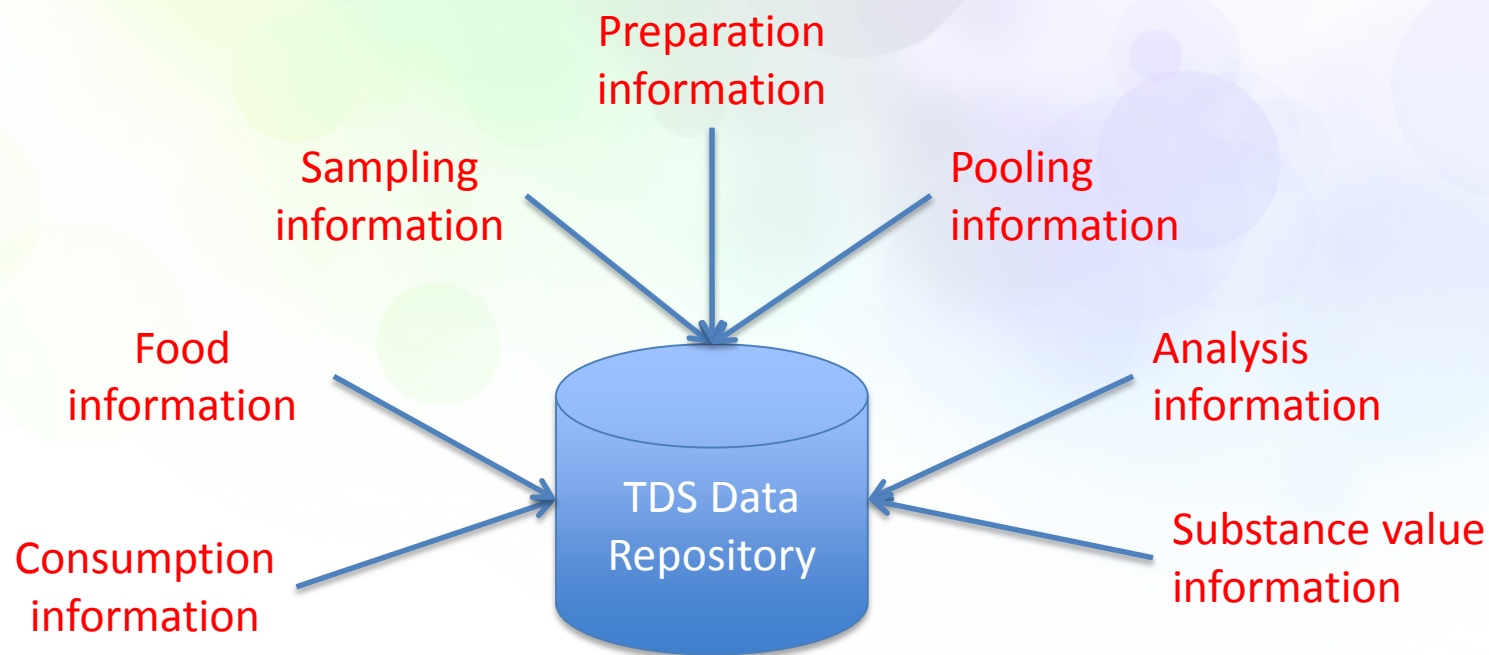
pool

Pooled food X

How does a TDS work?



Management of TDS Data



FC

TDS Pool value detail

2015, Helsinki, Finland

File

Save

Save & close

Close

Pool name

Apple, raw

Value id

2

Substance name

Selenium (Se)

Value

0.02

Value type

(MN) mean

Unit

(mg) milligram

Matrix unit

(TKG) per kg total food

Value

Lab analysis

Markers

Remarks

Files

Measurement limits

LOD

0.0002

LOQ

0.0003

LOR

0.0001

Statistics

Mean

0.02

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

Take Home Message II

Value documentation is
crucial for scientific
evidence.

What Data Needs to be Managed?

- ▶ Until now every country did its own TDS and collected data according to their requirements
- ▶ One goal in the TDS-Exposure project is to establish recommendations what data should be collected

What Data Amount Needs to be Managed?

The following numbers were identified:

- 200 foods
- 96 attributes per food and 49 for one substance value

Example:

200 food items with about 10 substances
-> at least 117'200 data points

Linkage of Food Items

Food Compilation Application

Food Composition

Food Name
Apple, fresh
Pear, fresh

Component	Value
Protein	0.3g/100g
Vitamin C	5mg/100g

Food Consumption

Person
Person1
Person2

Food	Amount
Apple, fresh	1 piece
Pear, fresh	1 piece

Total Diet Study

Food Name
Apple, fresh
Pear, fresh

Component	Value
Mercury	0mg/100g
Selenium	0mg/100g



Take Home Message III

Use a feasible tool to
manage your TDS and
make your data
sustainable.

How Data Needs to be Managed?

How to Work with a TDS Data Management System?

Assemble food list

Different data sets are needed:
food comp data, consumption
data and former TDS data

Assemble shopping list

EXCEL is a handy tool to
manipulate multiple entries.
Issue: 18 EXCEL files (one for
each food category)

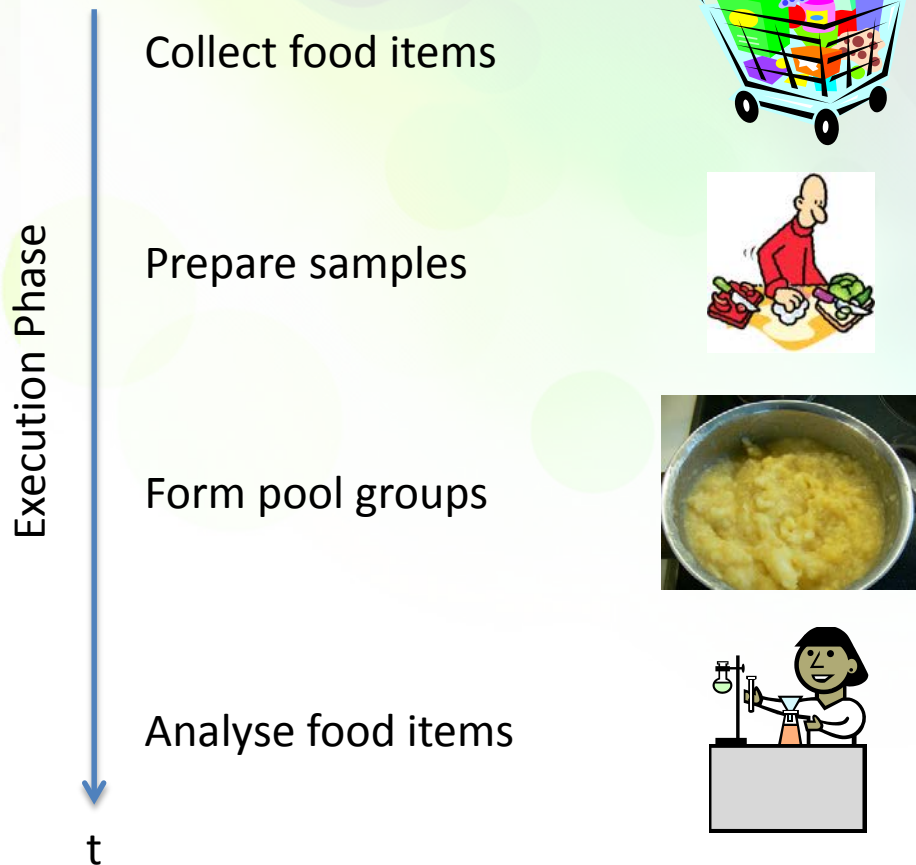
Form pool groups

Food Item	Energy	Protein	Carbohydrate	Fat	Fiber	Decision
101 Cereals	1000	10.0	70.0	10.0	0.0	Yes
102 Cereals	1000	10.0	70.0	10.0	0.0	Yes
103 Cereals	1000	10.0	70.0	10.0	0.0	Yes
104 Cereals	1000	10.0	70.0	10.0	0.0	Yes
105 Cereals	1000	10.0	70.0	10.0	0.0	Yes
106 Cereals	1000	10.0	70.0	10.0	0.0	Yes
107 Cereals	1000	10.0	70.0	10.0	0.0	Yes
108 Cereals	1000	10.0	70.0	10.0	0.0	Yes
109 Cereals	1000	10.0	70.0	10.0	0.0	Yes
110 Cereals	1000	10.0	70.0	10.0	0.0	Yes
111 Cereals	1000	10.0	70.0	10.0	0.0	Yes
112 Cereals	1000	10.0	70.0	10.0	0.0	Yes
113 Cereals	1000	10.0	70.0	10.0	0.0	Yes
114 Cereals	1000	10.0	70.0	10.0	0.0	Yes
115 Cereals	1000	10.0	70.0	10.0	0.0	Yes
116 Cereals	1000	10.0	70.0	10.0	0.0	Yes
117 Cereals	1000	10.0	70.0	10.0	0.0	Yes
118 Cereals	1000	10.0	70.0	10.0	0.0	Yes
119 Cereals	1000	10.0	70.0	10.0	0.0	Yes
120 Cereals	1000	10.0	70.0	10.0	0.0	Yes

Box 1	Sample characterization and number of subsamples	Box 2	Food items selection criteria
16	Sample name: ...	16	Food items selection criteria: ...
17	Items from the TDS food list: ...	17	Food items selection criteria: ...
18	Items from the TDS food list: ...	18	Food items selection criteria: ...
19	Items from the TDS food list: ...	19	Food items selection criteria: ...
20	Items from the TDS food list: ...	20	Food items selection criteria: ...
21	Items from the TDS food list: ...	21	Food items selection criteria: ...
22	Items from the TDS food list: ...	22	Food items selection criteria: ...
23	Items from the TDS food list: ...	23	Food items selection criteria: ...
24	Items from the TDS food list: ...	24	Food items selection criteria: ...
25	Items from the TDS food list: ...	25	Food items selection criteria: ...
26	Items from the TDS food list: ...	26	Food items selection criteria: ...
27	Items from the TDS food list: ...	27	Food items selection criteria: ...
28	Items from the TDS food list: ...	28	Food items selection criteria: ...
29	Items from the TDS food list: ...	29	Food items selection criteria: ...
30	Items from the TDS food list: ...	30	Food items selection criteria: ...



How to Work with a TDS Data Management System?



Only deviation from original plan must be documented.

How to Work with a TDS Data Management System?

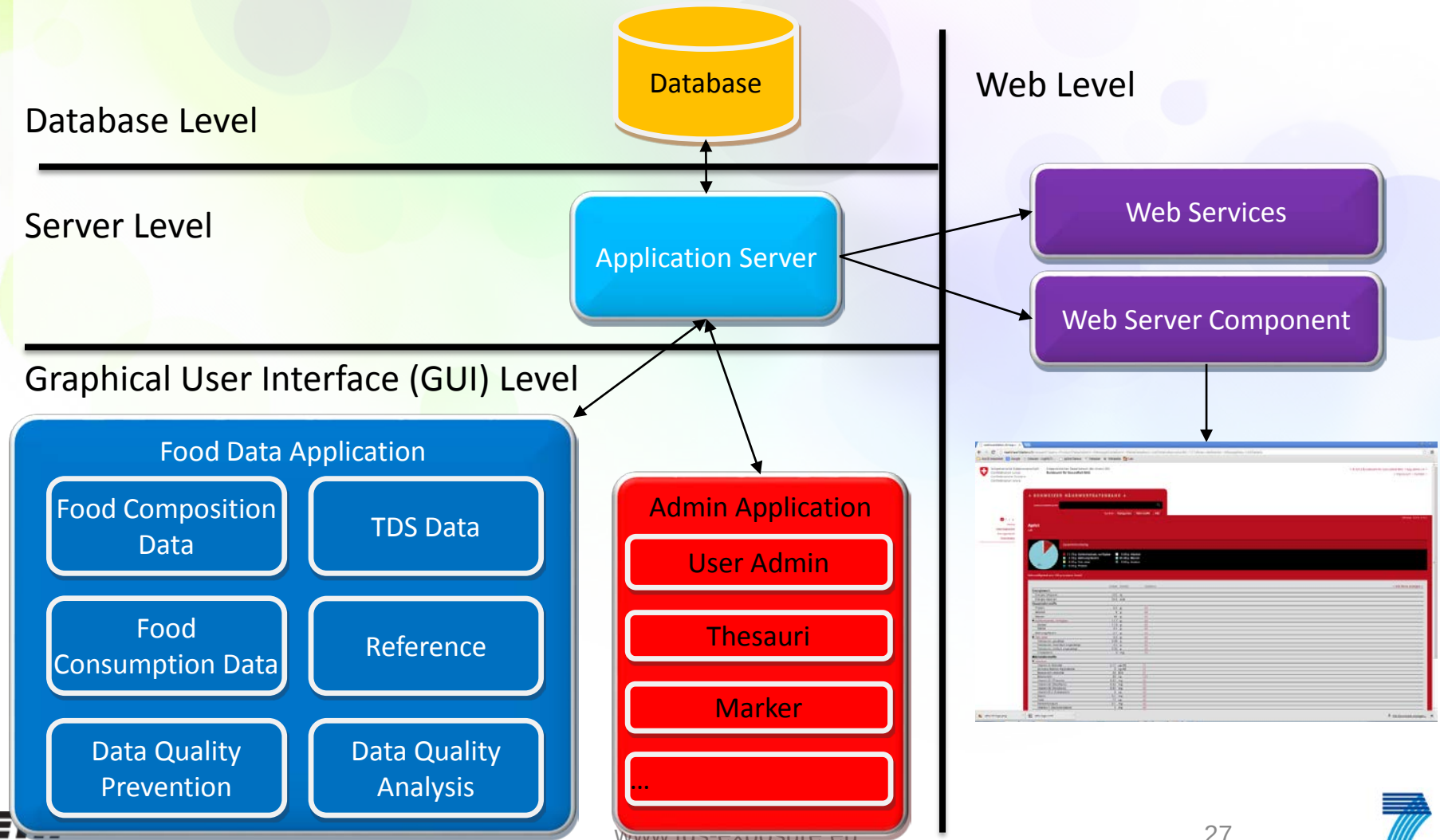


Take Home Message IV

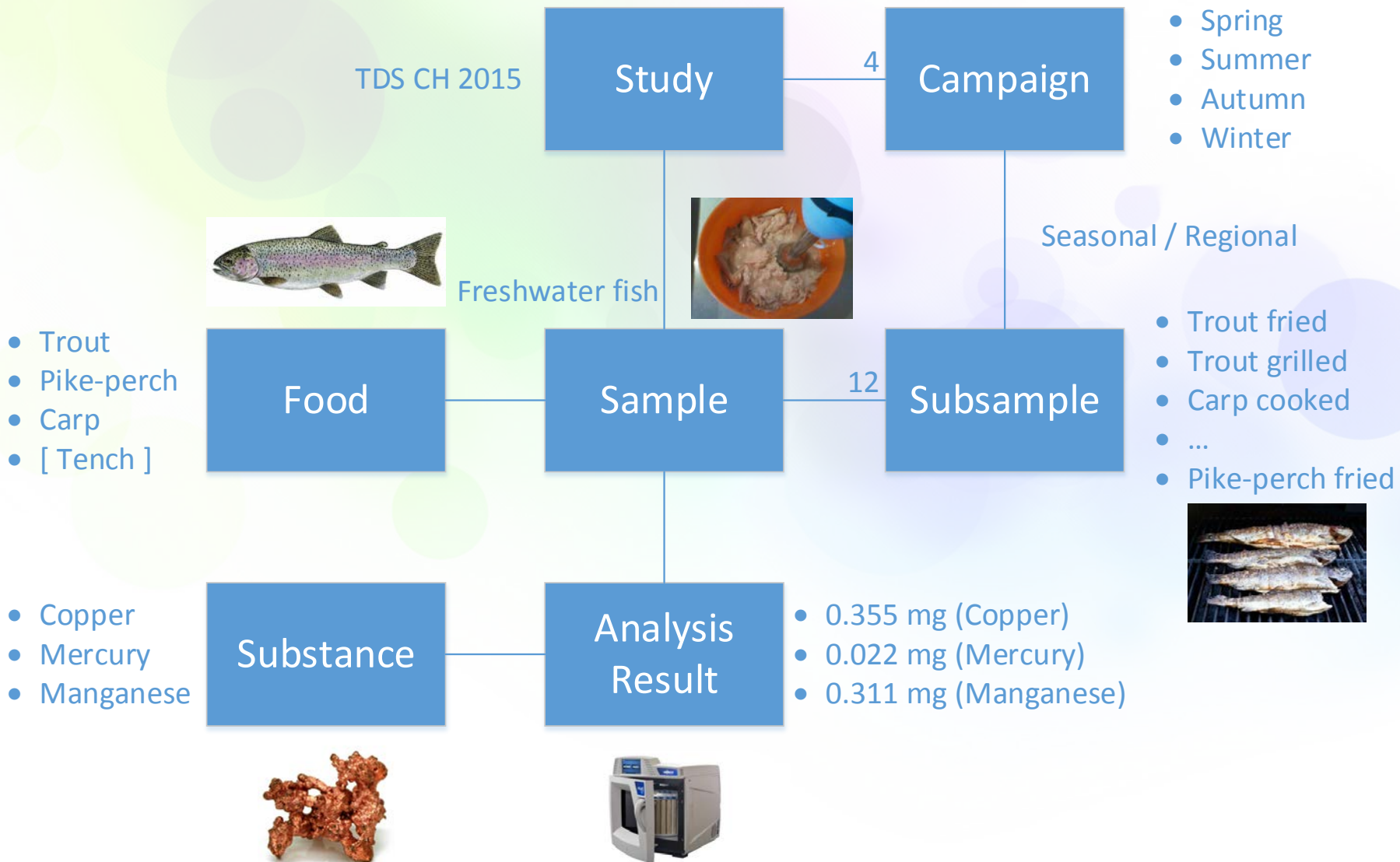
A good TDS data management system does not restrict you in your TDS approach. It rather supports your value documentation and risk assessment.

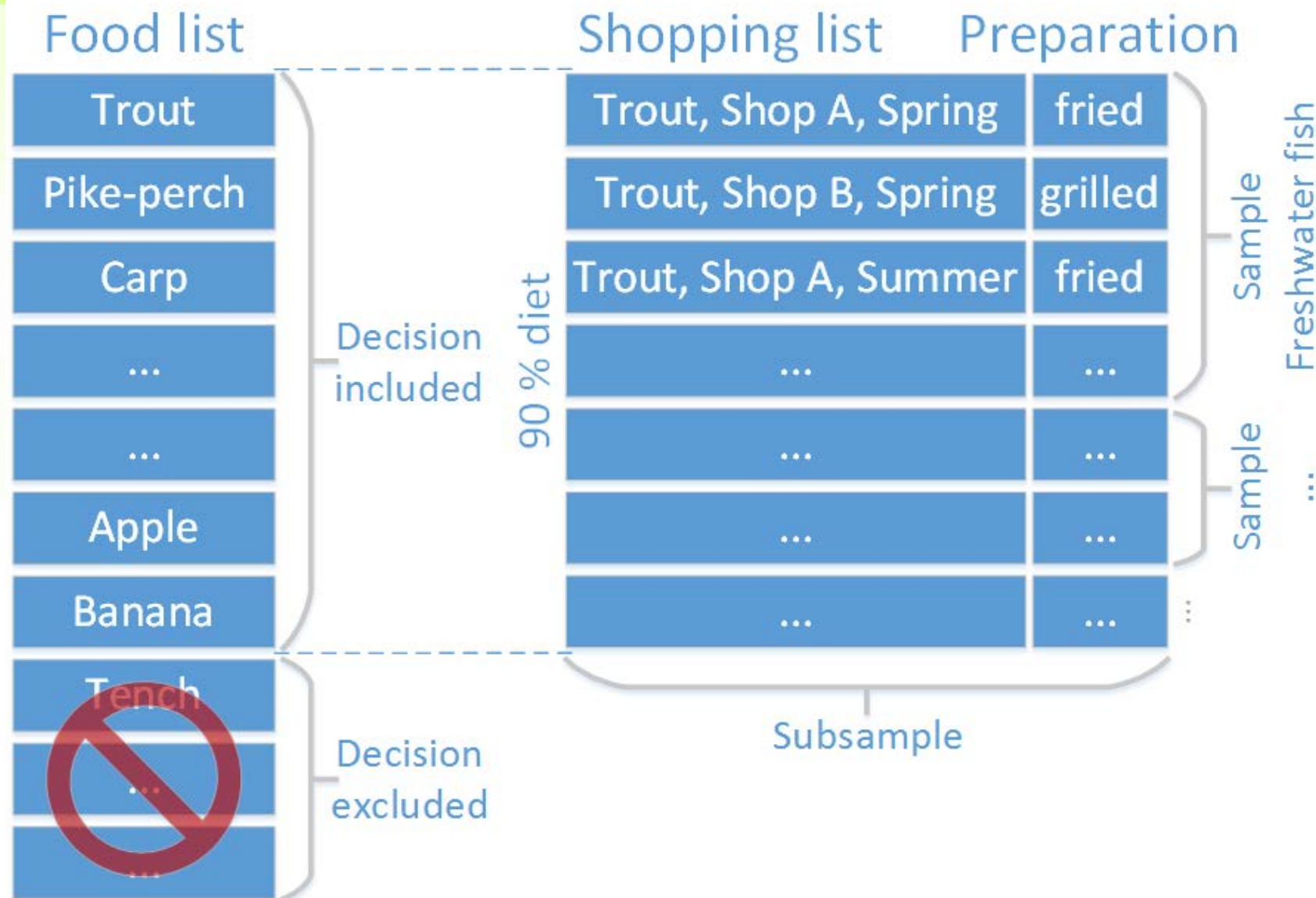
Architecture and Food Data Entities

Architecture of a Food Information System



TDS CH 2015





Use Cases in FoodCASE

Starting Mask - Overview

Search Area

Data Listing Area

- ▶ All or certain columns
- ▶ Content of substances
- ▶ Data markers
- ▶ LanguaL / FoodEx2 codes

Food / Sample items

Analytical results
(beneficial and harmful substances)

Playground v3.0.0 RC2

File View Tools Help

Single Values Aggregated Values Recipes References Versions TDS Foods TDS Pools TDS Studies

Food Search

Food criteria

Column: <all columns>

Function: contains

Term:

☐ Case sensitive ☐ Ignore time

Component criteria

Component:

< >

☒ And ☐ Or

< >

Search Clear

Component Search

Column: <all columns>

Function: contains

Term:

☐ Case sensitive ☐ Ignore time

Search Clear

Food Markers

Value Markers

LanguaL Codes

TDS Pools

ID	Pool Name	English Name	Sci...
1	Kaffeegetränke	Coffee beverages	
2	Cappuccino	coffee drink, cappuccino	
3	Teegetränke (schwarz, grün)	Tea beverages	
4	Kräutertee	Herbal and other non-tea infusions_herbal te...	
5	Früchtetee	Fruit infusion	
6	Robostee	Robos infusion	
7	Instant-(Früchte-)Tee-Getränke	Coffee, cocoa, tea and herbal drinks - instan...	
8	Pizza, Quiche, Flammkuchen	Pizza and pizza-like dishes	
9	gemischter Salat	Mixed vegetable salad	
10	Fertigsuppe Konserve	Soups (cooked, ready-to-eat)_preserve (can...	
11	Nudelsuppe	Soups (cooked, ready-to-eat)_noodle soup	
12	Gulasch, Frikassee, Geschnetzeltes	Meat-based dishes_goulash/fricassee	
13	Obstsalat	Fruit salad	
14	Kartoffelsuppe	Potato soup	
15	Gerichte mit Fleisch_Roulade	Meat based dishes_roulade	
16	Nudelsalat	Prepared pasta salad	
17	gefüllte Paprika, Kohlroulade	Dishes excluding pasta or rice dishes, sandwi...	
18	gemischtes Gemüse, Gemüsepfanne	Mixed vegetables, fried	
19	Fleischsuppe	Meat soup, with pieces	
20	Tomatensuppe	Tomato soup	

Refresh 147 rows Open New...

TDS Pool values: Fruit salad

ID	Substance Name	Code	Unit	Matrix Unit	Valu
37	Total copper	RF-00000168-CHE	mg	TKG	0.68
38	Total mercury	RF-00000170-CHE	mg	TKG	0
39	Total manganese	RF-00000177-CHE	mg	TKG	3.2

Refresh 3 rows Open New... Duplicate

Clipboard

Batch Aggregation

ID	Name
----	------

Refresh Batch process Clear

Managing Sample

- Mandatory fields
- Many other attributes for a sample
- Data quality panel

The screenshot shows a web application window titled "TDS pool detail, pool id: 6 / import source id: Import source". The interface includes a "File" menu with "Save", "Save & close", and "Close" options. Below this is a "TDS Pool" section with input fields for "Id" (6), "Pool name" (highlighted in orange), "English name" (Garlic), and "Scientific name" (No scientific name). To the right of these fields is a "Load picture..." button. A tabbed interface below shows "FoodEx2" as the active tab, displaying "FoodEx2 Exposure" (Garlic and similar (p) (A00GY)) and "FoodEx2 Facet" (Food Ex2 Facet). There are "Add..." and "Remove" buttons next to the facet list. A "FoodEx2 comment" field is also present. At the bottom, a "Data quality evaluation" section shows error messages: "The pool name cannot be empty." and "The pool name must have at least 2 characters." followed by "Current object quality is 59.523808%".

Managing Subsamples

- Attribute values depending on the table selection

TDS pool detail, pool id: 20 / import source id: 0527

File

Save Save & close Close

TDS Pool

Id 20

Pool name Tomatensuppe

English name Tomato soup

Scientific name

Study BFR pilot 2014 [id=1]

Load picture...

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

Id	Name	English name	Brand	Type	Origin	Place
186	Recipe				Recipe	
187	Recipe				Recipe	
188	Tomatensuppe (Standardrezeptur)		Fix & fertig, Tomaten Cremesuppe	Dose	D1	
189	Tomatensuppe mit Reis		Maggi, Tomatensuppe mit Reis	Maggi	Tüte	V1
190	Tomatensuppe (Standardrezeptur)		Knorr, Tomatensuppe	Knorr	Tüte	V1
191	Tomatensuppe (Standardrezeptur)		Knorr, Tomatensuppe Toscana	Knorr	Tüte	V1
192	Tomatensuppe (Standardrezeptur)		Erasco, medi. Tomatensuppe	Erasco	Dose	V1

New Remove

Sample info Preparation methods FoodEx2

Name Tomatensuppe (Standardrezeptur) English name Brand Fix & fertig, Tomaten Cremesuppe

Generic name Producer hergestellt für Norma: BUSS Fertig Campaign BFR campaign 2

Type Price Currency

Origin Dose Country of origin EAN code 20121488

Sampling place D1 Sampling date 2014-07-28 Preparation place

Unit size Number of units Packing type

Raw quantity Edible quantity Prepared quantity

Purchase minimum 1 Dose Sensorial test i.O. Date of expiry Jun 16, 2016

Pool amount Unit Sampled by

Remarks

Recipe procedure

Data quality evaluation

Managing Analysis Results

- ▶ Laboratory report
- ▶ Other scientific publication
- ▶ Basic idea:
Provide as much value documentation as possible
- ▶ Attach original published document

Managing Analysis Results

- Many available attributes
- Data quality panel

The screenshot shows a software window titled "TDS Pool value detail". It contains several input fields and a data quality evaluation section.

File menu: Save, Save & close, Close

Fields:

- Pool name: Obstsalat
- Substance name: Total copper
- Value: 0.68773036668
- Unit: (mg) milligram
- Value id: 37
- Value type: (MIN) mean
- Matrix unit: (TKG) per kg total food

Value tab (selected):

- Lab analysis: [empty]
- Markers: [empty]
- Remarks: [empty]
- Files: [empty]

Measurement limits:

- LOD: 0.04333333333333335
- LOQ: 0.13
- LOR: [empty]

Statistics:

- Mean: [empty]
- Median: [empty]
- Minimum: [empty]
- Standard deviation: [empty]
- Maximum: [empty]
- n: [empty]
- Uncertainty: 0.13924211552027702
- Maximum uncertainty: 0.13924211552027702

Date:

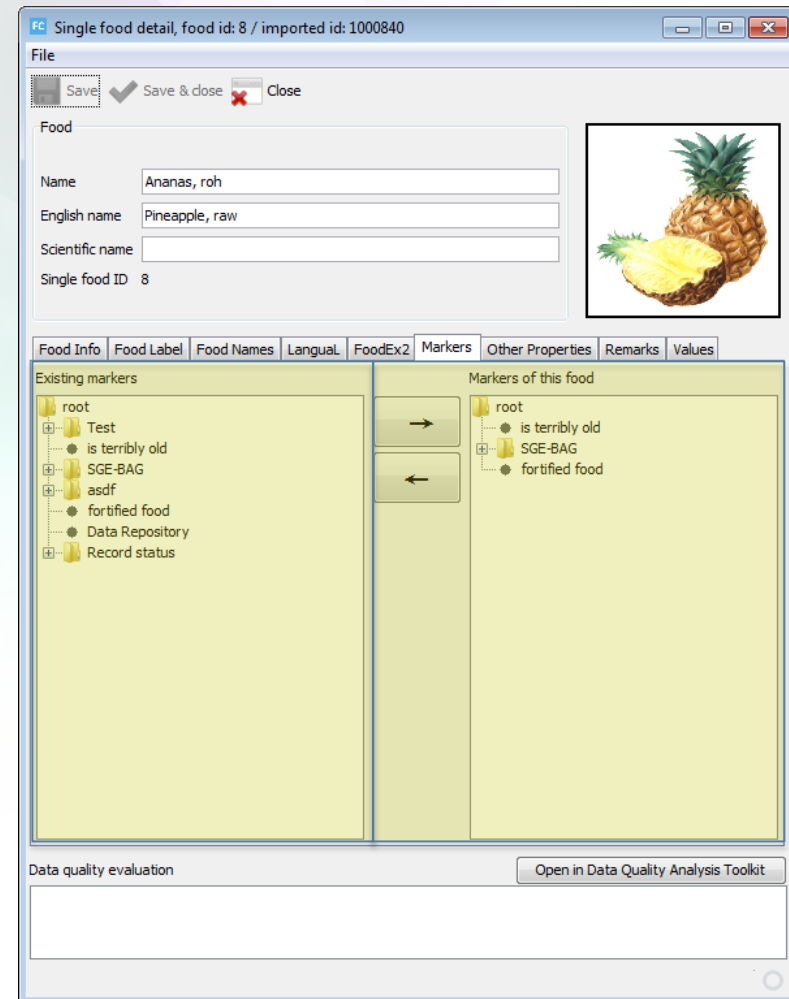
- Verifying date: [empty]
- Verifying by: [empty]

Data quality evaluation:

- Open in Data Quality Analysis Toolkit
- Verifying by field should not be empty.
- Current object quality is 92.30769%

Markers

- ▶ Available markers
- ▶ Assign or remove from current entity
- ▶ Definition of markers in the Admin client
- ▶ Custom naming
- ▶ Multilevel markers for TDS



Take Home Message V

FoodCASE contains all necessary concepts and its user interfaces are simple structured and intuitive.

Data Quality in FoodCASE

Data Quality (DQ)

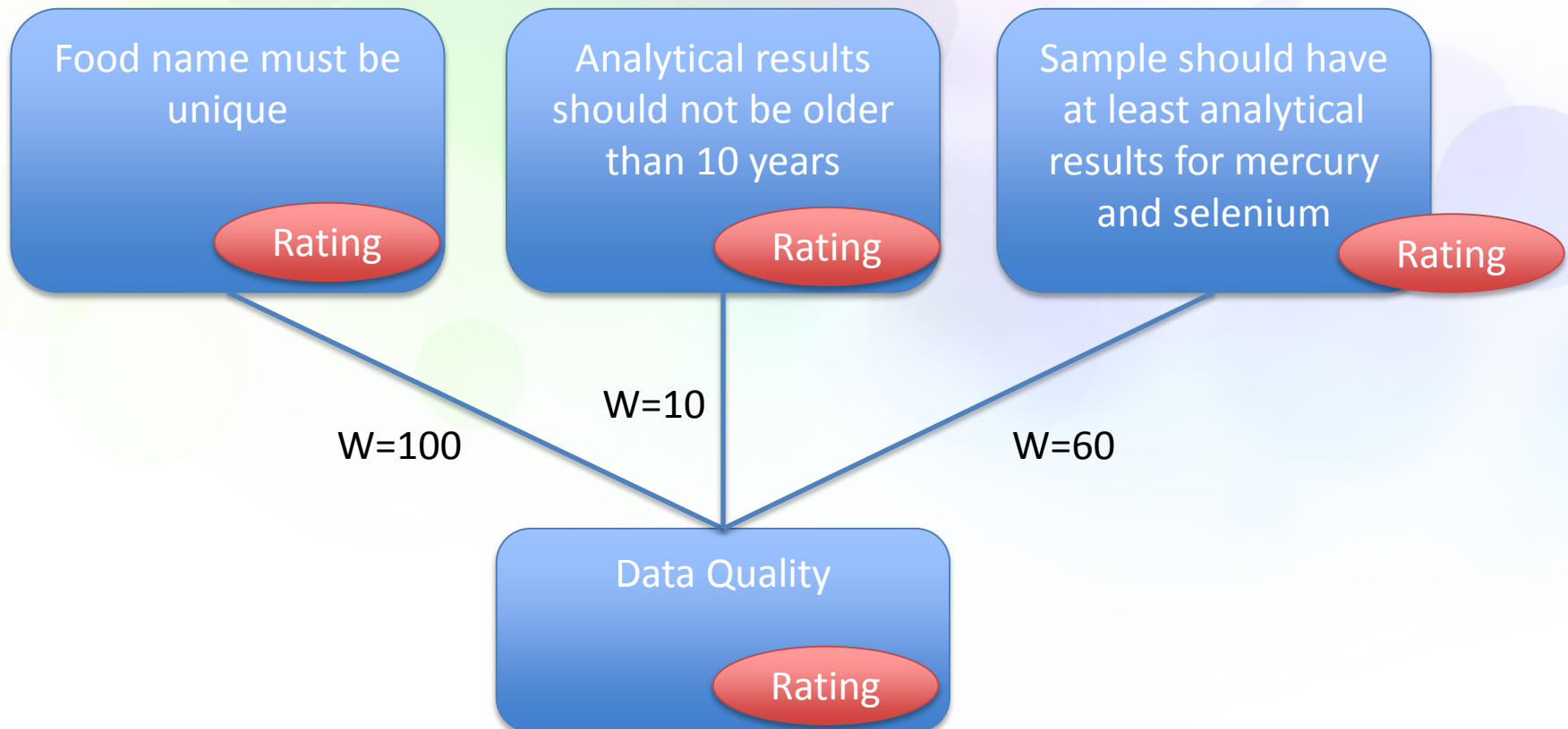
Definitions of data quality:

1. Data or information has quality if it is appropriate for the intended use
2. The quality of data or information depends on users' expectations and requirements

Data Quality Requirement (DQR)

- A data quality requirement is a concrete intrinsic or contextual constraint to data that helps to satisfy an aspect or construct of data quality.

Data Quality Requirements



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 Weights & Contracts

- ▶ User can define weights and threshold
 - 0: Not important, 100: Most important
 - Threshold between warning and error
 - Default weight
- ▶ A contract defines the weight (importance) for each of the constraints
- ▶ Default contract exists as fallback
- ▶ Users can define their (multiple) own contracts
- ▶ Contracts can easily be imported and shared

Types of DQ Requirements

Food name must be unique

Hard constraints → Error

Determines if data is deficient or not

Sample should have at least analytical results for mercury and selenium.

Soft constraints → Warning

The more requirements are satisfied, the better is data quality

Analytical results should not be older than 10 years

Indicators → Warning

Indicate better quality but quality improvement is not evident

Data Quality Rating

- ▶ Can be simple as 'satisfied' or 'not satisfied' or more complex, for instance 0 to 100
- ▶ Can be defined by user

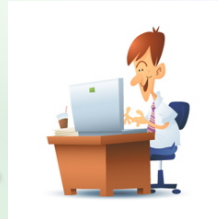
Food name must be
unique

Rating

DQF constraints & weighting - Example

- ▶ Sample: Kartoffelsuppe (engl. Potato soup)
- ▶ AnalysisResult: Total copper, 0.845 mg/kg
- ▶ Two constraints:
 - Sample.englishName.NotEmpty=80, active
(very important)
 - AnalysisResult.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 Quality Areas



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) (A000Y)

FoodEx2 Facet

Food Ex2 Facet

Add Remove

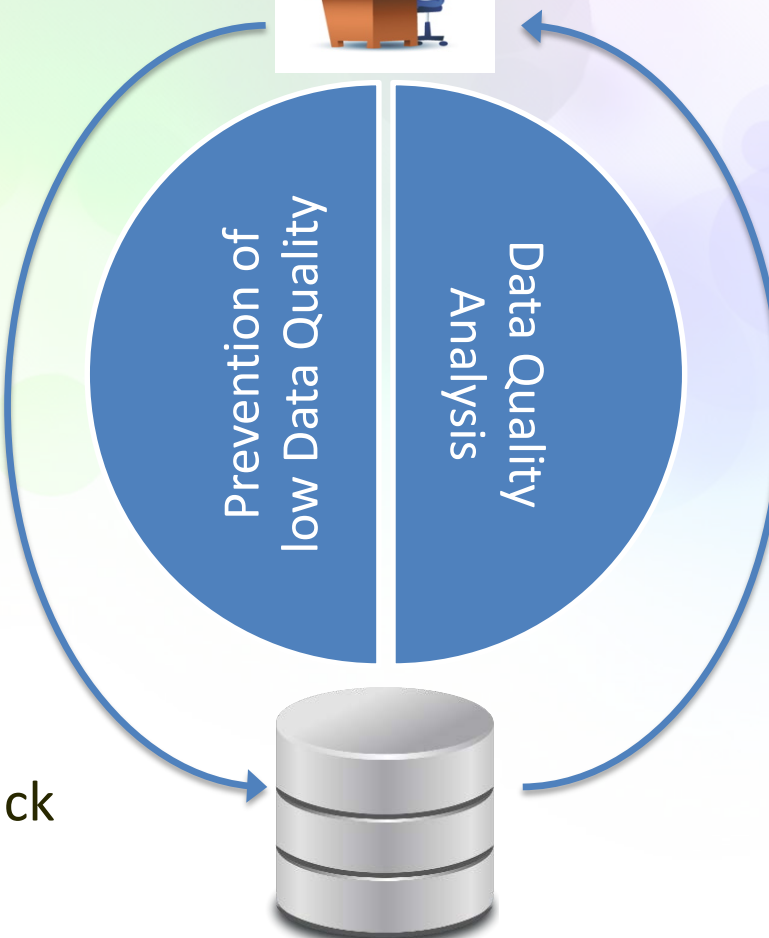
Comment

FoodEx2 comment

Data quality evaluation

The pool name cannot be empty.
The pool name must have at least 2 characters.
Current object quality is 59.523608%

Data quality input
validation & feedback



Data quality
analysis & feedback

Data quality administration

The screenshots show the FoodCASE Admin application interface, which is used for managing data quality constraints and validation settings. The application has a menu bar (File, Help) and a tabbed interface for different settings.

Top Screenshot (Entity and Group Selection): Shows the 'Entity and Group Selection' tab. It lists various entities and groups, including 'TdsPoolValue', 'TdsPoolValueChecked', 'CompletenessCheck', 'TdsPool', and 'All Ungrouped Constraints'. The 'Show warnings during input validation' checkbox is checked.

Middle Screenshot (General Tds Validation Settings): Shows the 'General Tds Validation Settings' tab. It includes a checkbox for 'Override constraint contract weight if its payload attribute is set', a 'Default constraint weight' slider set to 80, and a 'Warning to error threshold' slider set to 40. Below these are informational text blocks about quality metrics and scales.

Bottom Screenshot (Edit Contract): Shows the 'Edit Contract' tab. It displays a table of constraints with columns for Entity Name, Entity Field (optional), Constraint Name, Can be over..., Weight, and Is Inactive. The table lists various constraints like 'ValidMinMax', 'Size', 'Min', 'Min', 'NotEmpty', 'PoolFatConsistent', 'Min', 'UniquePoolName', 'ValidRangeValue', and 'NotEmpty'.

Entity Name	Entity Field (optional)	Constraint Name	Can be over...	Weight	Is Inactive
TdsPoolValue		ValidMinMax	☐	80	☐
TdsPool	name	Size	☐	90	☐
TdsPoolValue	minimum	Min	☐	100	☐
TdsPoolValue	value	Min	☒	70	☐
TdsPoolValue	verifyingBy	NotEmpty	☒	20	☐
TdsPool	englishName	NotEmpty	☒	30	☐
TdsPool		PoolFatConsistent	☐	90	☐
TdsPoolValue	maximum	Min	☐	30	☐
TdsPool		UniquePoolName	☐	100	☐
TdsPoolValue		ValidRangeValue	☐	80	☐
TdsPool	name	NotEmpty	☒	100	☐

Data quality input validation

TDS Pool value detail

File

Save Save & close Close

Pool name Schokoladenkuchen Value id 7

Substance name Mercury and derivatives

Value Value type (MN) mean

Unit (mg) milligram Matrix unit (TKG) per kg total food

Value Lab analysis Markers Remarks Files

Measurement limits

LOD 0.0002 LOQ 0.0003

LOR 0.0001

Statistics

Mean 0.01 Median 1

Minimum 0.02 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

The pool value should be between minimum and maximum values.
 The pool value must be greater than or equal to 0.
Current quality indicator: 59 quality points (out of 100)
 'Verifying by' field should not be empty.

Data quality analysis

Run Tds Data Quality Analysis

Analysis Configuration

Selected Contract:

Entity and Group Selection

- ☒ All Analyzable Tds Entities and Groups
 - ☒ TdsPoolValue
 - ☒ CompletenessChecks
 - ☒ All Ungrouped Constraints
 - ☒ TdsPoolValueChecks
 - ☒ TdsPool
 - ☒ All Ungrouped Constraints

Analysis Results Overview ☒ Violations per Entity Pie Chart ☐ Violations per Entity Bar Chart ☐ Constraints per Weight Range Chart

Analysis

Analysis Results

```

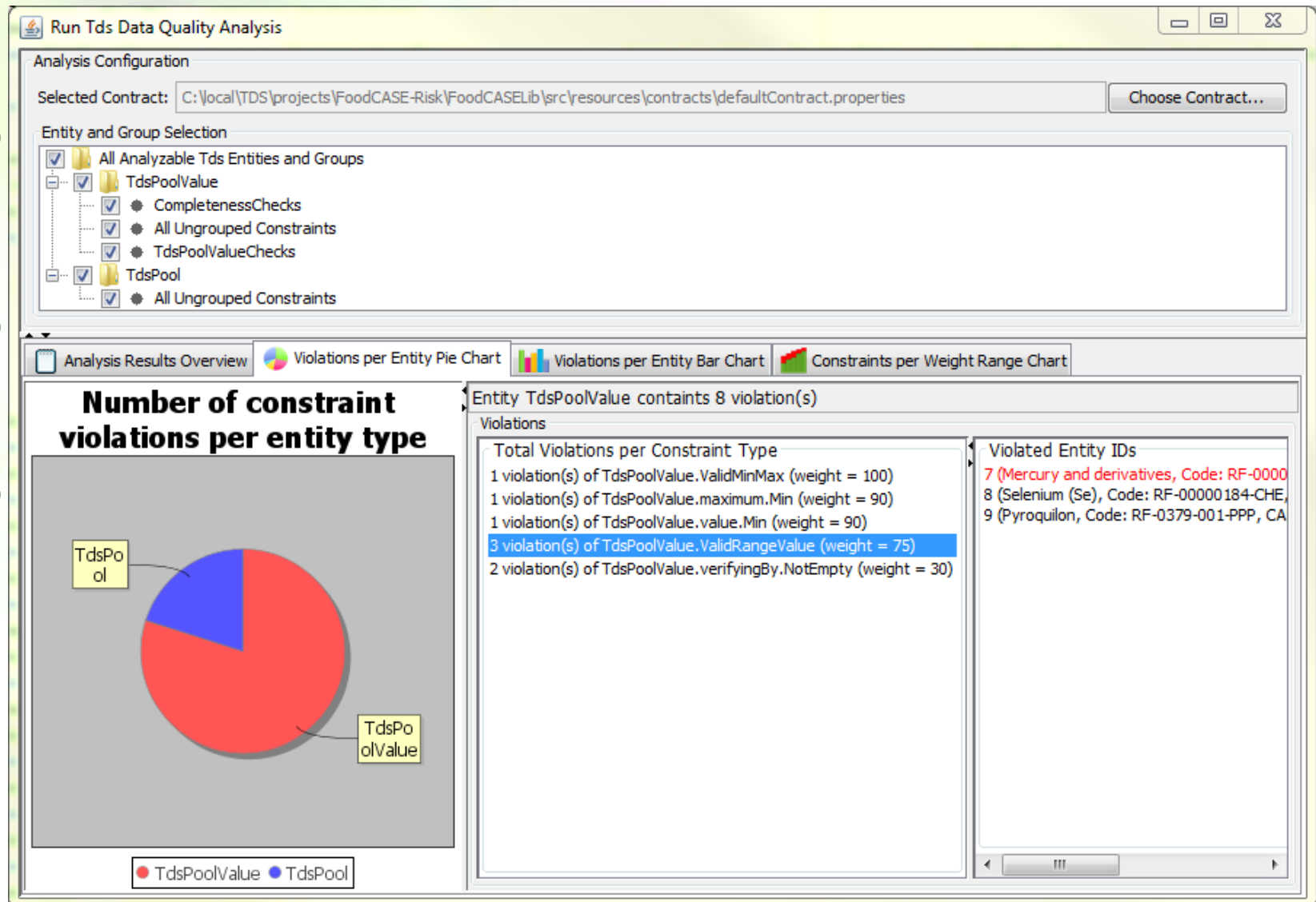
12 objects of type TdsPoolValue were validated and 8 were violation-free
72 TdsPoolValue constraints were validated and 8 violations were found

8 objects of type TdsPool were validated and 6 were violation-free
40 TdsPool constraints were validated and 2 violations were found

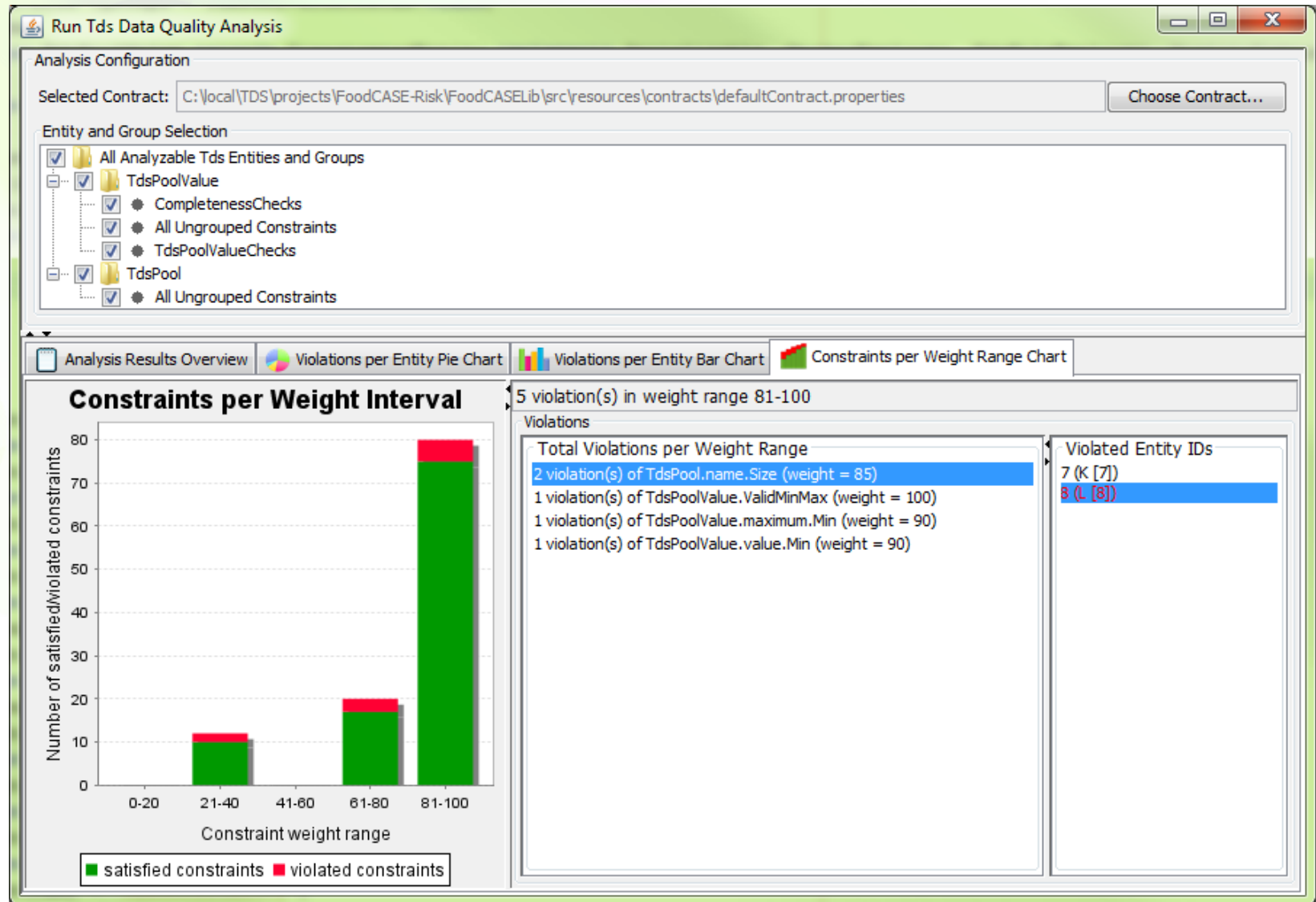
SUMMARY OF RESULTS
-----
12 objects of type TdsPoolValue with 90 quality points (out of 100)
8 objects of type TdsPool with 95 quality points (out of 100)

The overall quality indicator of the validated data is 92 quality points (out of 100)
    
```

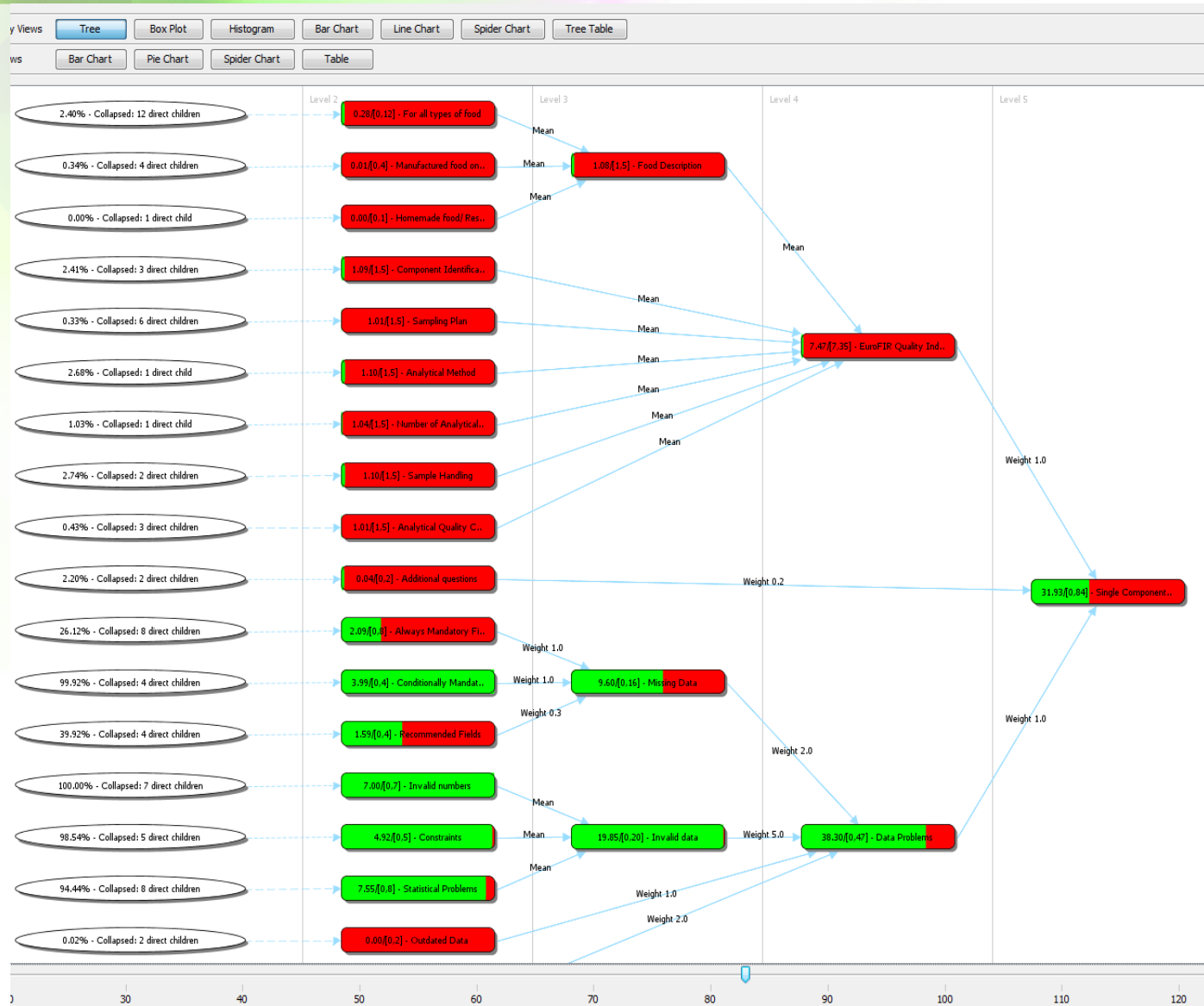
Data quality analysis



Data quality analysis



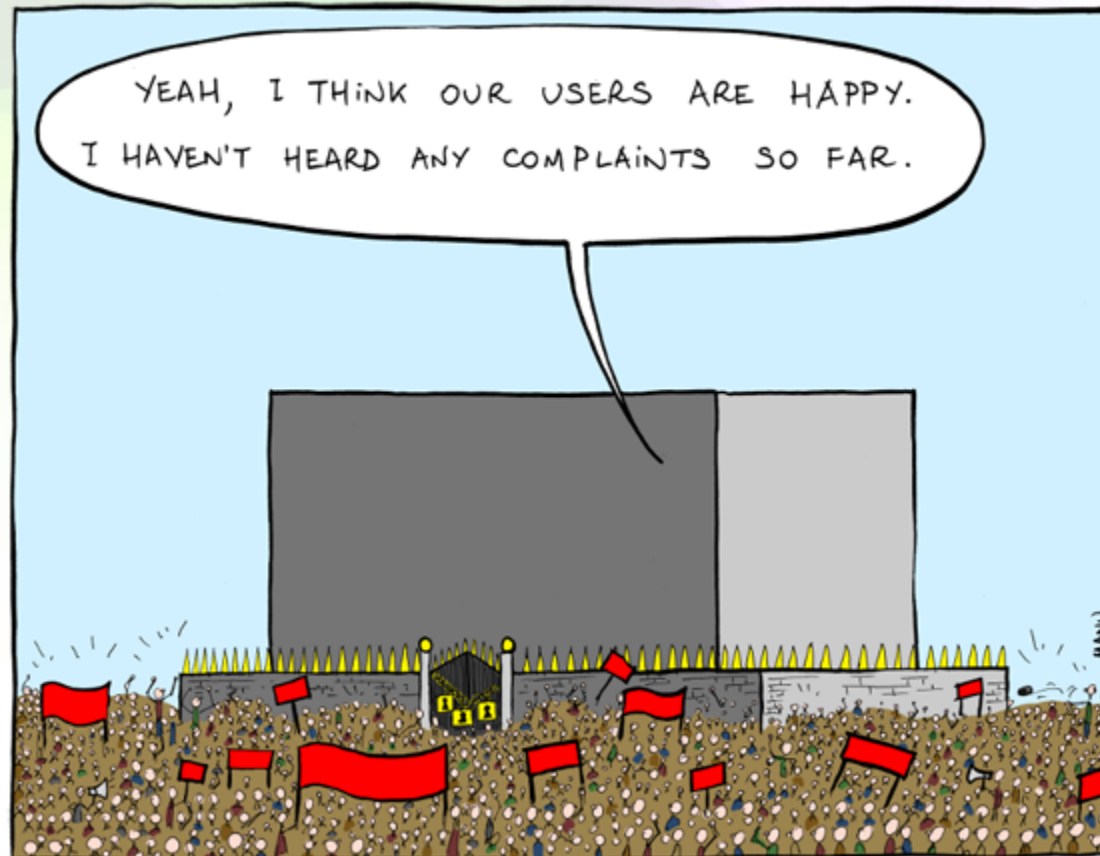
Data quality analysis



Take Home Message VI

Data quality requirements can differ dependent on the utilisation of data. Thus, a flexible feedback is necessary.

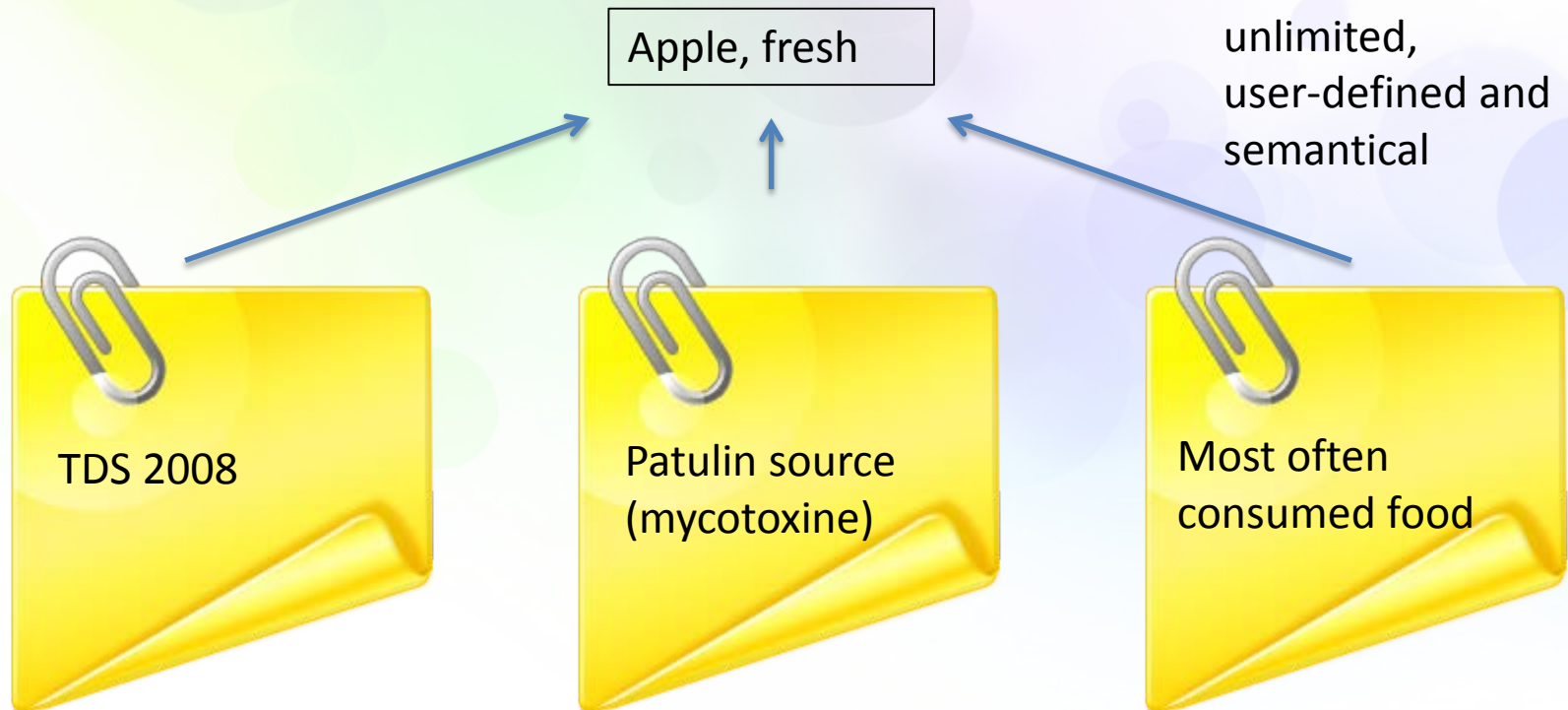
Your feedback / discussion



Thank you for your attention

Concepts for Data Management

TDS Data Markers



TDS Data Markers

TDS



Fortified



Concept 2:
Structure markers to
simplify overview

TDS Data Markers

Search food items which were not
used in the TDS 2008
but in TDS 2010
and are fortified with vitamin C

Concept 3:
Markers are searchable

“TDS 2008” AND NOT “TDS 2010” AND “fortified with vit C”

Take Home Message

Markers is a flexible concept for semantical data tagging.

Reports and Exports of TDS Data

Reporting/Exporting means

- to search and prepare existing data,
- to bring it into a usable form such as table or graphic,
- and to put it into a **human or computer readable** file format such as PDF or EXCEL

Reports and Exports of TDS Data

Challenge: Reports are useful but exist in endless number of variations and implementation costs.

Solution:

- Data must be selectable for use in report/export
- User should be able to define their own reports in terms of form and format

Take Home Message

Report and export functionalities should satisfy needs and in best case can be extended by users.

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 not empty') , 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

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