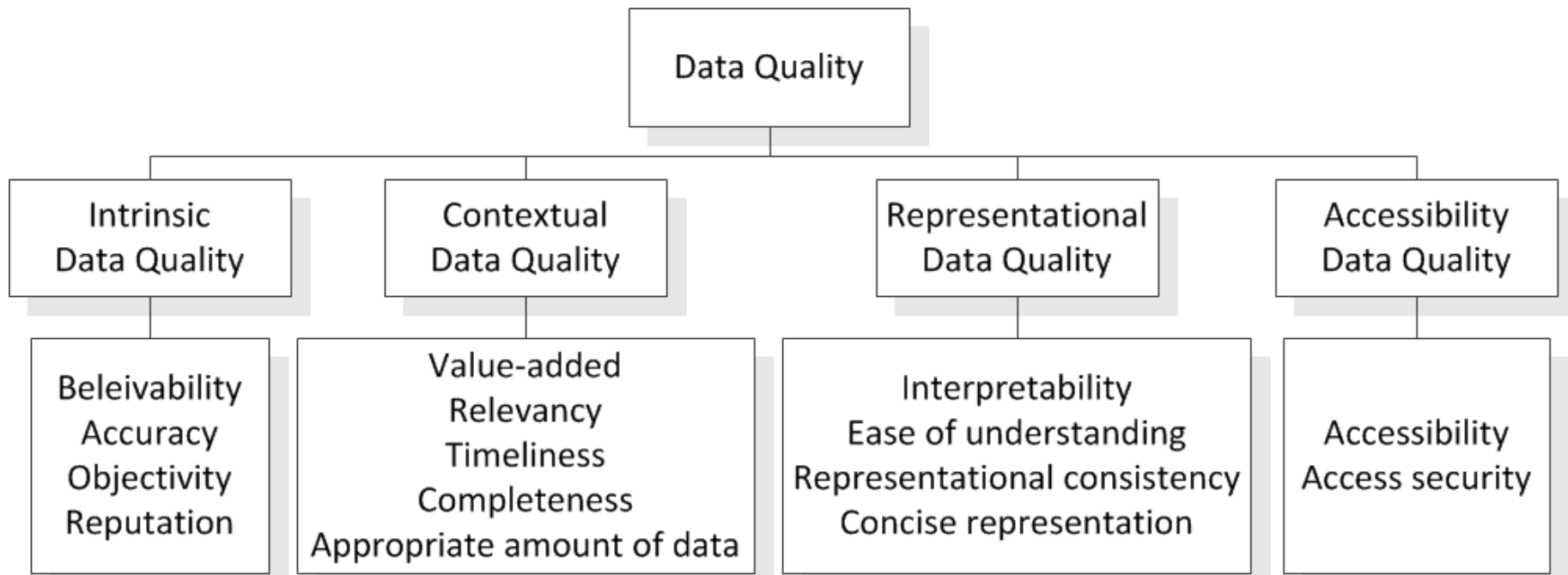


How to Give Feedback on Data Quality: A Study in the Food Sciences

David Weber, Karl Presser, & Moira Norrie (ETH Zurich)

Data Quality (DQ)

- Fitness for use
- Critical issue in scientific databases
- Data quality (DQ) dimension for the various aspects
- DQ framework by Wang and Strong
- Lack of general tools, mechanisms and metrics



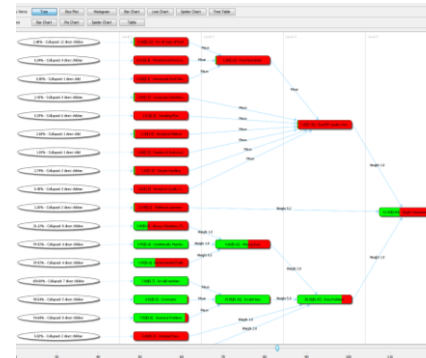
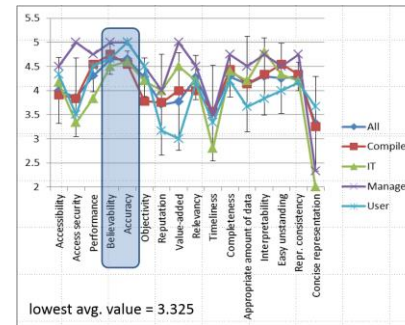
Overview

Long-term goal: general DQ framework

→ Visual access to DQ metrics

→ Indicators of actions

- First step:
Detailed study of DQ requirements
(ECIS 2014)
- Second step:
How to give feedback on DQ to users of an application
(ECIS 2015)
 - Approach
 - DQ analysis toolkit
 - Survey



Motivation

ValidationDemo 16:08

Name 

Participants  # of participants

Date Please, enter the number of participants

Type Work

Email android@inmite.

Full name

David Smith

✓ Name looks great.

Email address

david.smith@mydomain

✓ We will email you a confirmation.

Create a password

.....



✓ Password could be more secure.

Choose your username

david

✗ This username is already taken!

Suggestions: david45519112 · david07165106 · david18196435

Name

David

Smith

Choose your username

david.smith

@gmail.com

Someone already has that username. Note that we ignore periods and capitalization in usernames. Try another?

Available: ds4559913

Create a password

You can't leave this empty.



There was a problem with your request

Your email address has an invalid email address format. Please correct and try again.

Registration

New to Amazon.com? Register Below.

My name is: David Smith

My e-mail address is: david.smith@mydomain

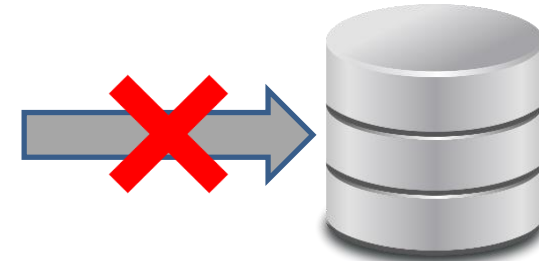
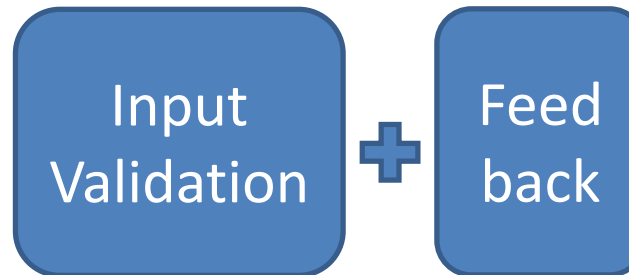
Motivation cont.

- Valid / invalid not sufficient
- Distinguish several levels of DQ
- Requires another concept for validation and metrics for DQ
- How and when to give feedback

How to ensure DQ in an information system?

Data input process

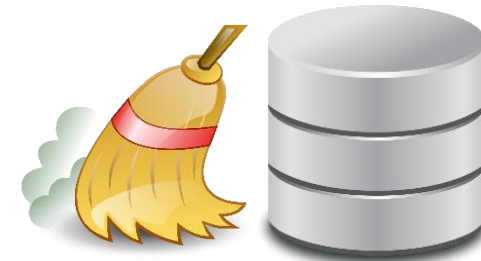
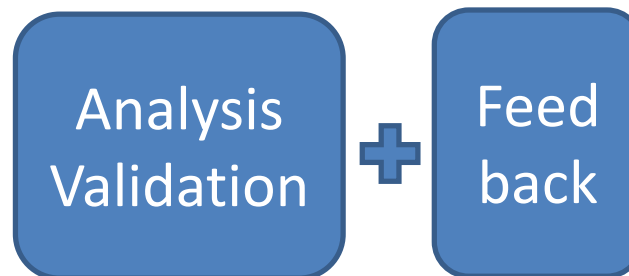
- User Interface
- Business logic
- Database



→ Restriction:
Data of poor quality
cannot be entered

Data analysis

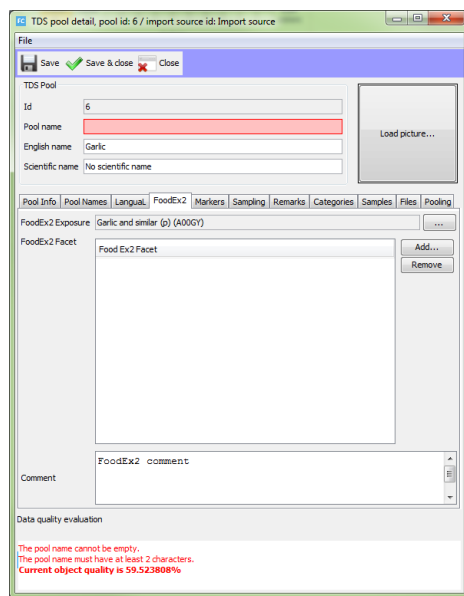
- Database



→ Data cleaning

Data Quality Input & Analysis Feedback (DQIF / DQAF)

DQIF



TD5 pool detail, pool id: 6 / import source id: Import source

File Save Save & close Close

TD5 Pool

Id: 6

Pool name:

English name: Garlic

Scientific name: No scientific name

Load picture...

Pool Info | Pool Names | Languages | FoodEx2 | Markers | Sampling | Remarks | Categories | Samples | Files | Pooling

FoodEx2 Exposure: Garlic and similar (p) (A00G)

FoodEx2 Facet:

Add... Remove

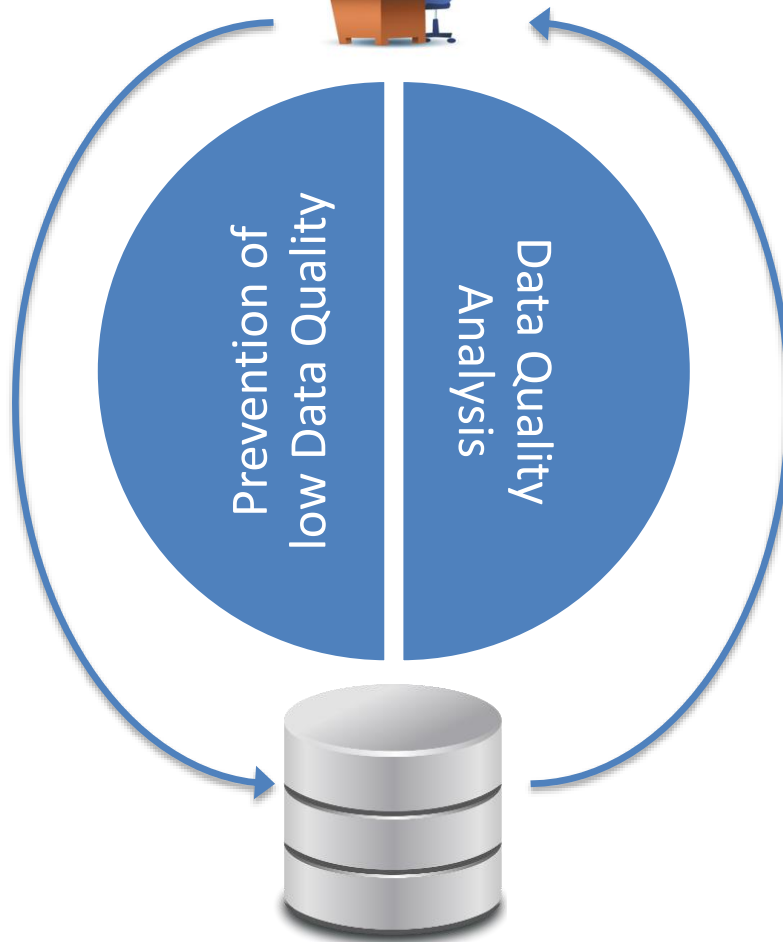
FoodEx2 comment:

Comment:

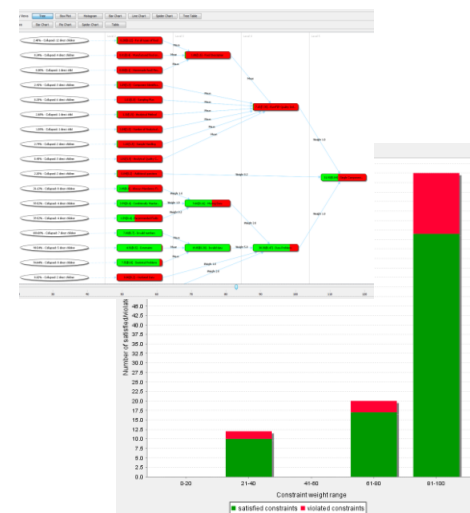
Data quality evaluation

The pool name cannot be empty.
The pool name must have at least 2 characters.
Current object quality is 59.523809%

Data quality input
validation
& direct feedback on
individual entities



DQAF



Data quality analysis
& feedback on
multiple entities

General DQ Framework

Concept

- Enhanced functionality for defining, validating, measuring and visualising DQ
- Helps users monitor constraints and DQ
- Includes the validation and feedback for both DQIF and DQAF

DQ Requirements

- 156 identified specific to the food science domain
- Definition of own DQ measurements
- Configuration of customised DQ feedback

Approach

Central element: DQ Requirements (DQRs)

- Hard constraint
 - Necessary condition
 - Data quality is deficient
 - Errors
- Soft constraints & indicators
 - Validated in a more relaxed way
 - Recommendation
 - Increase or decrease data quality
 - Warnings

DQIF Example

Aggregated food detail, food id: 293[head [1] (public)] / imported id: 852008

File

Save Save & close Close

Food

Food id 293

Name Beeren Crunch (bio-familia AG)

English name

Scientific name

Version head [1] (public) Public ☒ Marked for recomputation ☐

Beeren Crunch

Own Categories	Language	FoodEx2	Markers	Other Properties	Remarks	Values	Aggregation
Food Info	Food Label	Conversion Factors	Food Names				

Manufactured prepackaged food

Brand name Producer

Commercial name

Generic name

Classification

EuroFIR classification

Retention factor classification

Eurocode2 classification

Data quality evaluation

Open in Data Quality Analysis Toolkit

English name: Mandatory field empty

Retention factor classification: Without a classification, all retention factors are assumed to be 1.

If this food is used for recipe calculation, then density can be required.

DQAF Run

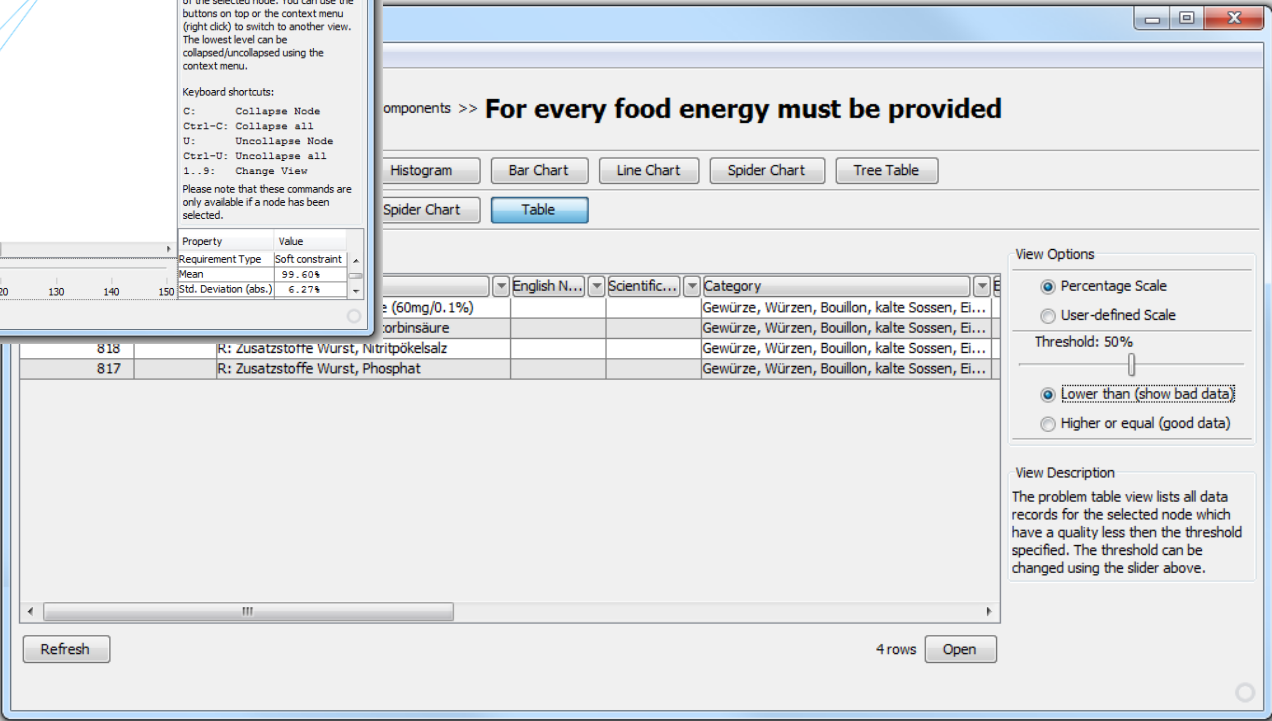
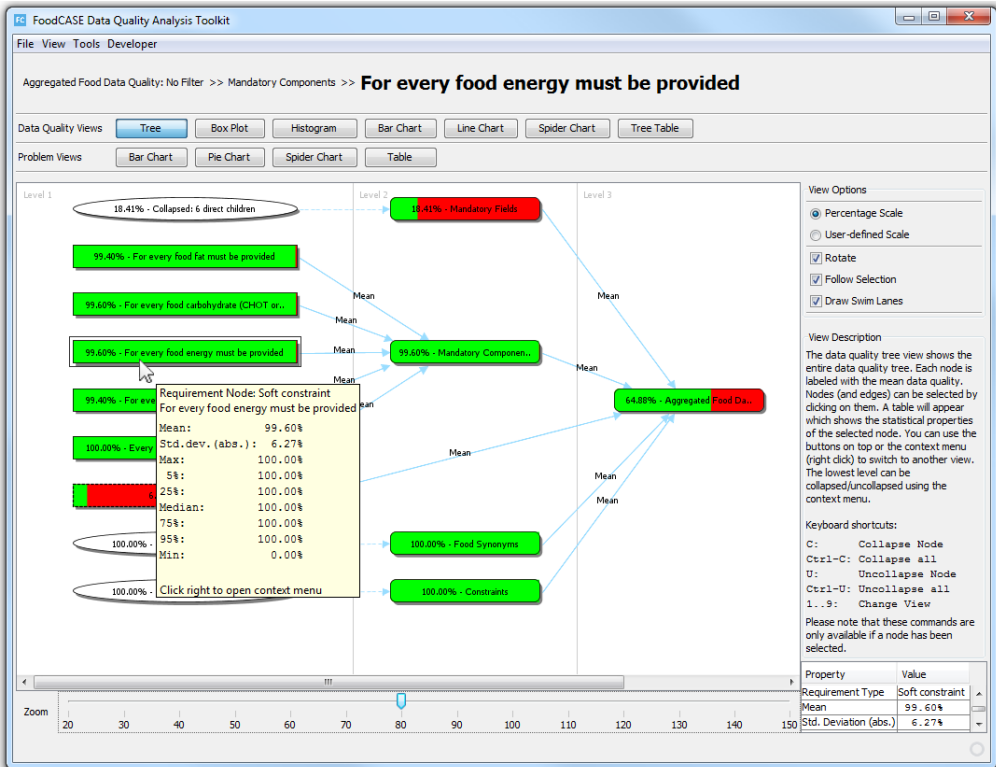
- Run a DQ assessment
 - All relevant data is fetched from the database
 - All DQRs are checked
 - Presentation of the DQ is prepared
 - The results are stored in the database for later analysis
- Triggered manually or by a timer
- Variety of different charts and tables support the user in judging where action is needed

DQAF DQ Measurement - Example

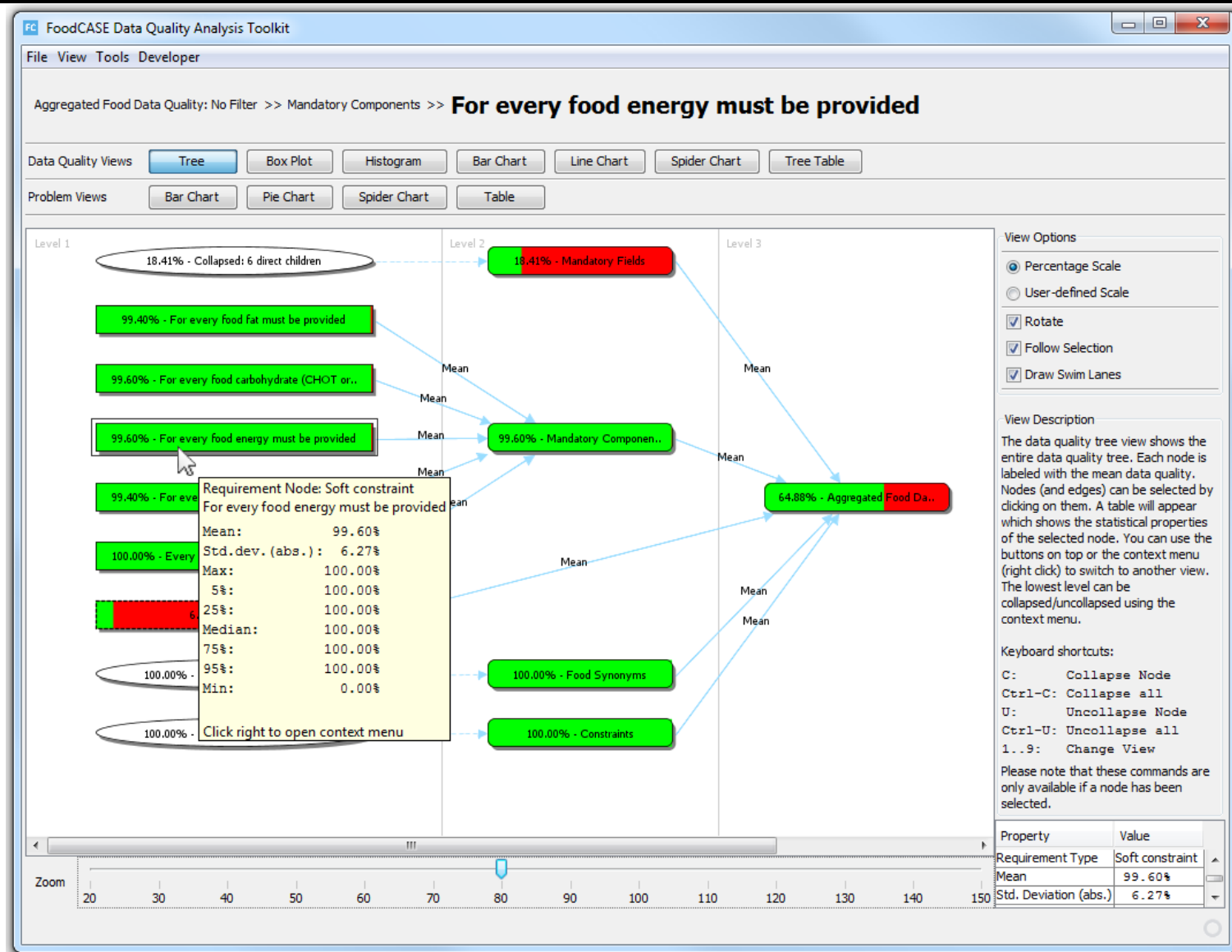
Example DQR for customers:

- 'For every person, last name and first name must be provided'
- Last name but not the first name is known
- DQR is only partially fulfilled
- DQ could be defined to be 0.7 (70%)

DQAF: DQ Views & Problem Views



DQAF: Tree Views



DQAF: DQ Views & Problem Views

The screenshot shows the FoodCASE Data Quality Analysis Toolkit interface. The title bar reads 'FoodCASE Data Quality Analysis Toolkit'. The menu bar includes 'File', 'View', 'Tools', and 'Developer'. The main content area displays the path 'Aggregated Food Data Quality: No Filter >> Mandatory Components >> For every food energy must be provided'. Below this, there are two rows of view selection buttons: 'Data Quality Views' (Tree, Box Plot, Histogram, Bar Chart, Line Chart, Spider Chart, Tree Table) and 'Problem Views' (Bar Chart, Pie Chart, Spider Chart, Table). The 'Table' button is selected. The main data area, titled 'Aggregated foods', contains a table with 5 columns: 'Aggregated...', 'Recipe...', 'Name', 'English N...', 'Scientific...', and 'Category'. It displays 4 rows of data. To the right of the table is a 'View Options' panel with radio buttons for 'Percentage Scale' (selected) and 'User-defined Scale', a 'Threshold: 50%' slider, and radio buttons for 'Lower than (show bad data)' (selected) and 'Higher or equal (good data)'. Below this is a 'View Description' box. At the bottom left is a 'Refresh' button, and at the bottom right are '4 rows' and an 'Open' button.

FoodCASE Data Quality Analysis Toolkit

File View Tools Developer

Aggregated Food Data Quality: No Filter >> Mandatory Components >> **For every food energy must be provided**

Data Quality Views: Tree, Box Plot, Histogram, Bar Chart, Line Chart, Spider Chart, Tree Table

Problem Views: Bar Chart, Pie Chart, Spider Chart, **Table**

Aggregated foods

Aggregated...	Recipe...	Name	English N...	Scientific...	Category
1126		R: Vitamin C, Ascorbinsäure (60mg/0.1%)			Gewürze, Würzen, Bouillon, kalte Sossen, Ei...
816		R: Zusatzstoffe Wurst, Ascorbinsäure			Gewürze, Würzen, Bouillon, kalte Sossen, Ei...
818		R: Zusatzstoffe Wurst, Nitritpökelsalz			Gewürze, Würzen, Bouillon, kalte Sossen, Ei...
817		R: Zusatzstoffe Wurst, Phosphat			Gewürze, Würzen, Bouillon, kalte Sossen, Ei...

View Options

- ☒ Percentage Scale
- ☐ User-defined Scale

Threshold: 50%

☒ Lower than (show bad data)

☐ Higher or equal (good data)

View Description

The problem table view lists all data records for the selected node which have a quality less then the threshold specified. The threshold can be changed using the slider above.

Refresh 4 rows Open

Evaluation

- Presentation of our information system with the integrated DQ analysis toolkit to food science experts
- Subsequently conducted a survey
- Participants had a variety of roles
- 16 participants completed the questionnaire

Evaluation cont.

49 questions on various topics:

- DQ in general
- Who should be able to define DQ requirements
- How should DQ requirements be defined
- How should DQ feedback be given to users
- Distinction between hard constraints (HC) and soft constraints (SC)
- Personal information and user skills

Results and Findings

1. An information system should clearly offer the functionality to analyse and improve DQ.
2. Access to data should be controlled and restricted. At the same time, providing users with the possibility to recall the change history of data is clearly important where changes should be traceable to users.
3. It should be possible for the institution using the software to define DQ requirements, but this should be done by a non-IT person and limited to 'power users'.
4. DQ requirements should be managed in one place and in a unified way.

Results and Findings cont.

5. It should be possible to rate the relevancy of each DQ requirement because not all requirements have the same importance.
6. It is clearly not preferable to provide the possibility to deactivate all DQ requirements.
7. A summary such as '5 of 9 DQ requirements are satisfied' would motivate users to increase DQ.
8. A panel at the bottom of a GUI is a good solution for the DQ feedback but, for a lot of users, it would be preferable if DQ feedback was displayed close to the corresponding field(s).

Results and Findings cont.

9. An accentuation such as highlighting or colouring the corresponding field(s) and the usage of colours to indicate different types of DQ requirements is considered helpful.
10. The distinction between HCs and SCs is useful and both types should exist.
11. The number of HCs should be as small as possible with most users preferring to have more SCs than HCs.

Questions



Conclusions

Our information system FoodCASE already provides the following

- The functionality to analyse and improve DQ.
- The access to data is controlled and restricted.
- It is possible to recall the change history and each change of data is traceable to a given user.
- DQRs can be defined by certain power users.
- It is possible to rate the relevancy of each DQR by defining custom DQ feedback trees with custom weights of certain DQRs.
- The definition of DQRs as hard constraints, soft constraints or indicators is possible and the user is free to decide how to categorise individual DQRs.
- Feedback distinction between different types of DQRs is visualised in the DQ evaluation panel at the bottom of each screen.

DQIF Aspects

- Where
 - Close to the input data
 - Outside of the input area (above or below)
- How
 - Textual
 - Symbols
 - Colouring (messages, field labels or field itself)
- When
 - single data item is entered
 - all data in an input area has been entered

DQIF Relevant Information

Usually feedback about one data entity

- An error has occurred
- Where the error occurred
- How the error can be repaired
(includes type of error)

Usually feedback about multiple or all entities in a database at the same time

- Where are corrections required
→ Visualisations
- Navigation: Overview → Individual data entities

DQAF DQ Measurement

- DQ measures rendering dependent on DQR type (HC / SC & Ind.)
- DQR has to map every data record in the DQ entity to a DQ value
- DQ value between 0.0 (worst) and 1.0 (best)
- Once all DQRs are gathered, similar DQRs can be grouped together. This process can be repeated recursively ending up with a DQ tree

Example DQR for customers:

- ‘For every person, last name and first name must be provided’
- Last name but not the first name is known
- DQR is only partially fulfilled
- DQ could be defined to be 0.7 (70%)

Future Work

Based on the evaluation outcome, we plan to

- Define and manage the DQRs in one place in a unified way, probably by changing the DQR definitions to annotations in the entities and performing the validation by the means of BeanValidation (Java Specification Request 349).
- Provide a DQ summary in each screen such as '5 of 9 DQRs are satisfied' in order to motivate users to increase DQ.
- Keep our evaluation panel for detailed feedback but add short DQ feedback close to the appropriate fields, probably by an additional accompanied symbol such as a light bulb with attached tooltip.

DQ Analysis Toolkit – Start Screen

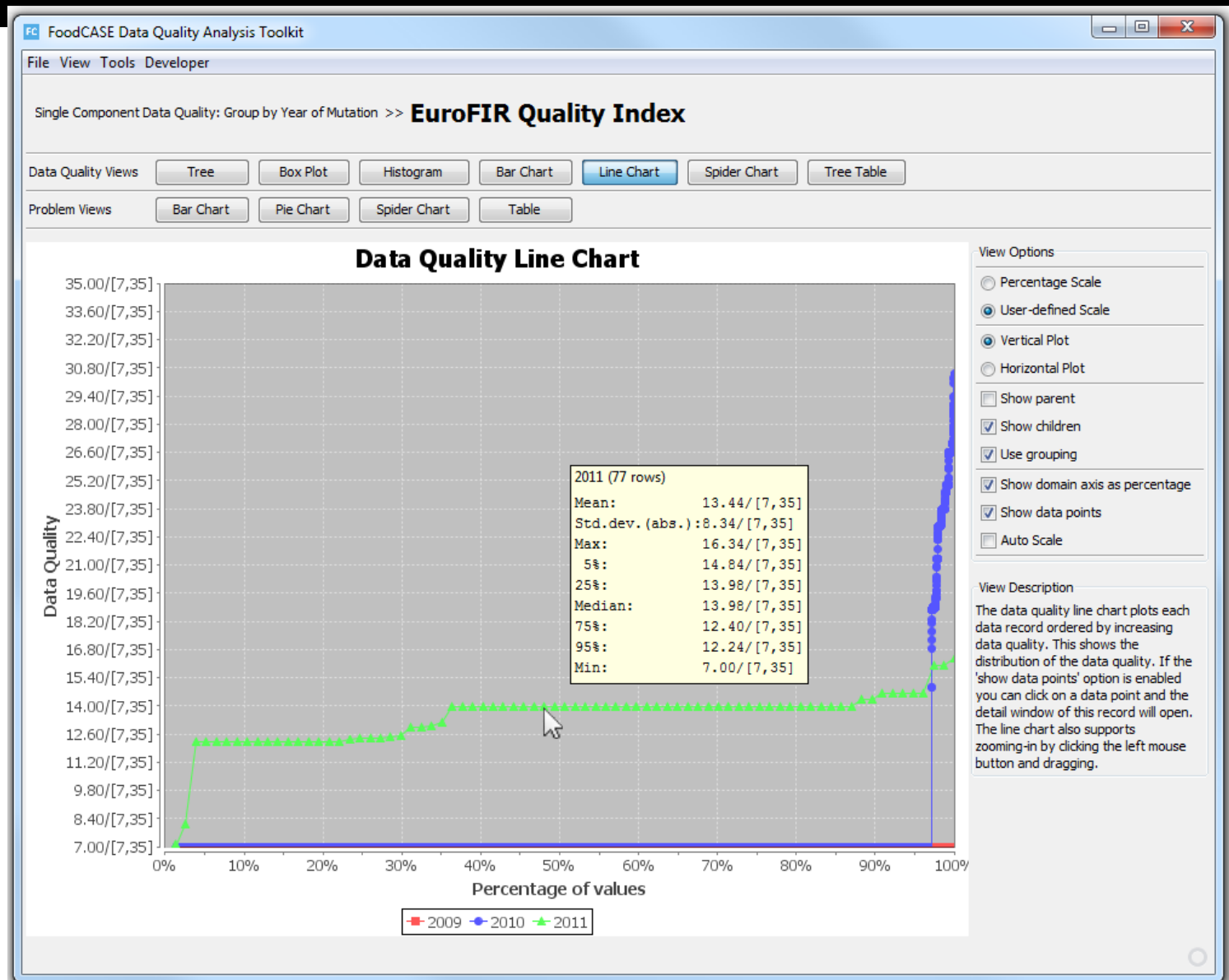
FoodCASE Data Quality Analysis Toolkit

File View Tools Developer

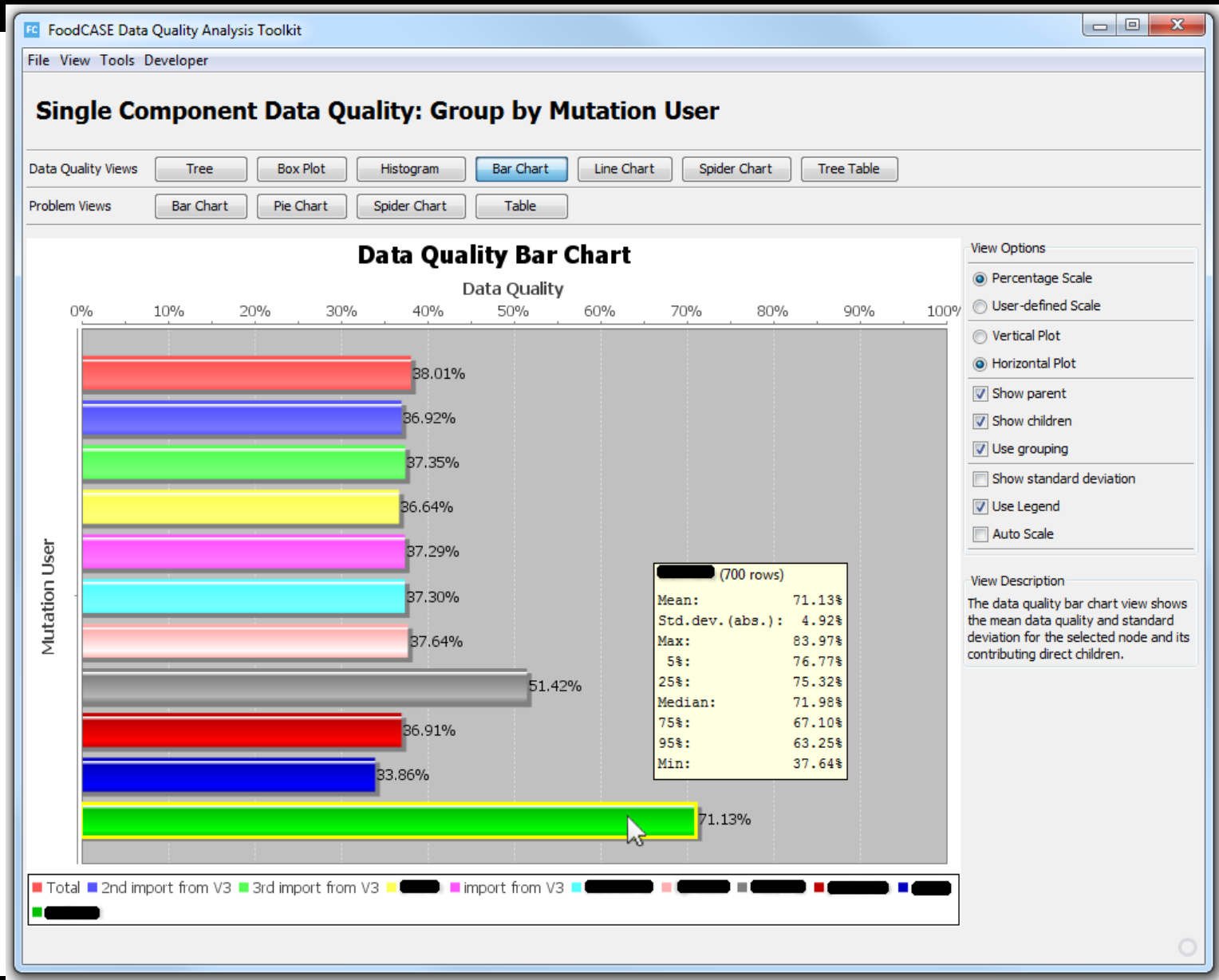
Welcome to the FoodCASE Data Quality Analysis Toolkit

1. Please select a data quality analysis tree definition
Aggregated Food Data Quality (rmock, 2011-05-21 21:51:03.539)
2. Specify filter or grouping criteria if desired
No Filter <no parameter>
3. Select a quality assessment
78 (2011-08-05 17:33:30.568, Created by Timer)
4. Treatment of not applicable requirements
Ignore
5. Select a presentation style
Data Quality Tree
6. Gather data and do calculation
Go

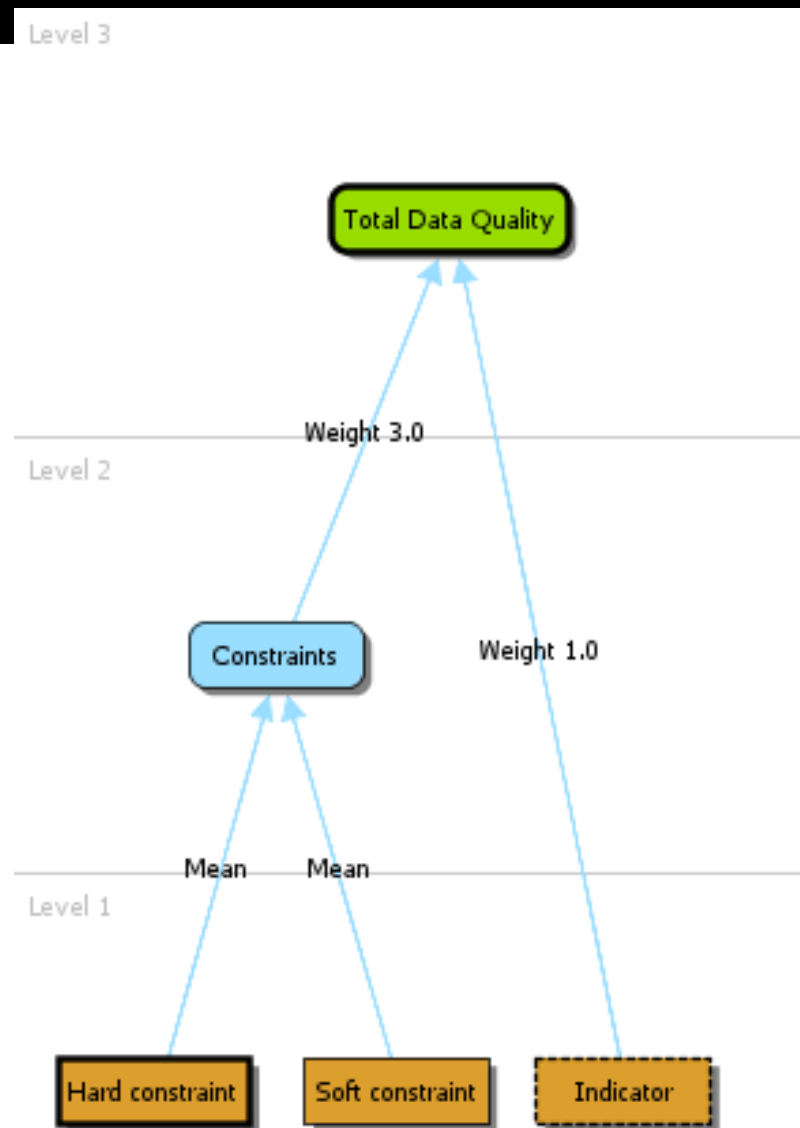
DQ Analysis Toolkit – Line Chart



DQ Analysis Toolkit – Bar Chart



DQ Analysis Toolkit – Simple Tree Definition



DQ Analysis Toolkit – Aggregation Types

- Mean
- Median
- Minimum
- Maximum
- Geometric Mean
- Weighted Mean

DQ Analysis Toolkit – Tree Editor

FoodCASE Data Quality Analysis Toolkit

File View Tools Developer

Tree Editor: Single Food Data Quality (rmock, 2011-05-23 21:21:59.101)

Use **Drag & Drop** to add new nodes to the tree

New Node

- Aggregation Node
- Requirement Node
 - Indicator
 - No more indicators
 - Soft constraint
 - No more soft constraints
 - Hard constraint
 - No more hard constraints

Level 1

- Mustfield EuroFIR classification
- Mustfield English food name
- Scientific name or brand name must exist
- Mustfield restaurant or homemade
- For homemade food recipe description must be prov..
- Mustfield retention factor classification
- Collapsed: 5 direct children
- Fill Factor
- Collapsed: 4 direct children
- Collapsed: 2 direct children

Level 2

- Mandatory Fields
- Mandatory Components
- Food Synonyms
- Constraints

Level 3

- Single Food Data Quality

Editor Options

- ☒ Rotate
- ☒ Follow Selection
- ☒ Draw Swim Lanes
- ☒ Use Tooltips

Properties of selected node or edge

Property	Value
Node Name	Mandatory Components
Node Type	Aggregation Node
Node Id	238
Level	2
Index	2
User Scale Min	
User Scale Max	
Aggregation Type	Mean

Help

The tree editor lets you define your data quality tree. Use drag & drop from the panel on the left hand side, the context menu or the keyboard shortcuts to edit it. You can add new requirement and aggregation nodes. New edges can be created by clicking on an aggregation node (keep the left mouse button pressed) and dragging. The table above shows the properties of the selected node or edge. Those properties printed in bold face can be edited. For example it is possible to specify the aggregation type of an aggregation node. Possible options include:

- Mean
- Median
- Minimum
- Maximum
- Geometric Mean
- Weighted Mean

Keyboard shortcuts:

- R: Add Requirement Node
- A: Add Aggregation Node
- Del: Delete Node/ Edge
- C: Collapse Node (Levels 1+2 only)
- Ctrl-C: Collapse all
- U: Uncollapse Node (Levels 1+2 only)
- Ctrl-U: Uncollapse all

Aggregation Node Constraints
Type: Mean
Click left and drag to create new edges
Click right to open context menu

Zoom: 20 30 40 50 60 70 80 90 100 110 120 130 140 150

DQ Analysis Toolkit – Filtering and Grouping

- Filter by Id (only consider a single data record)
- Filter/group by creation user
- Filter/group by year of creation
- Filter/group by mutation user
- Filter/group by year of mutation

DQR definition

- DQIF: Custom Java validation classes
- DQAF: SQL statements

DQAF Data Quality Requirements (DQR) View

FoodCASE Admin

File Help

User administration Import [not implemented] EuroFIR Thesauri User Thesauri Versioning Food Indexer (LanguaL) **Data Quality Analysis**

Data Quality Analysis Toolkit Administration

Quality Entities Quality Entity Filter **Quality Requirements** Quality Assessments Run Quality Assessment Quality Assessment Timer Maintenance

Quality Requirements

Entity	Name	Creation	C...	Mutation	Mut...
154	Aggregated Compone... SP: Mean <= Max	Jun 13, 2011 8:28:56 PM	rmock	Jun 24, 2011 3:25:35 PM	rmock
155	Aggregated Compone... SP: Mean >= Min	Jun 13, 2011 8:29:58 PM	rmock	Jun 24, 2011 3:25:44 PM	rmock
156	Aggregated Compone... SP: Median <= Max	Jun 13, 2011 8:31:09 PM	rmock	Jun 24, 2011 3:25:51 PM	rmock
157	Aggregated Compone... SP: Median >= Min	Jun 13, 2011 8:33:21 PM	rmock	Jun 24, 2011 3:25:58 PM	rmock

Quality Requirement Detail

Save Save & close Cancel

Requirement Id: 155

Entity: Aggregated Component

Name: SP: Mean >= Min

Description: Mean must be >= minimum value

Type: Hard constraint

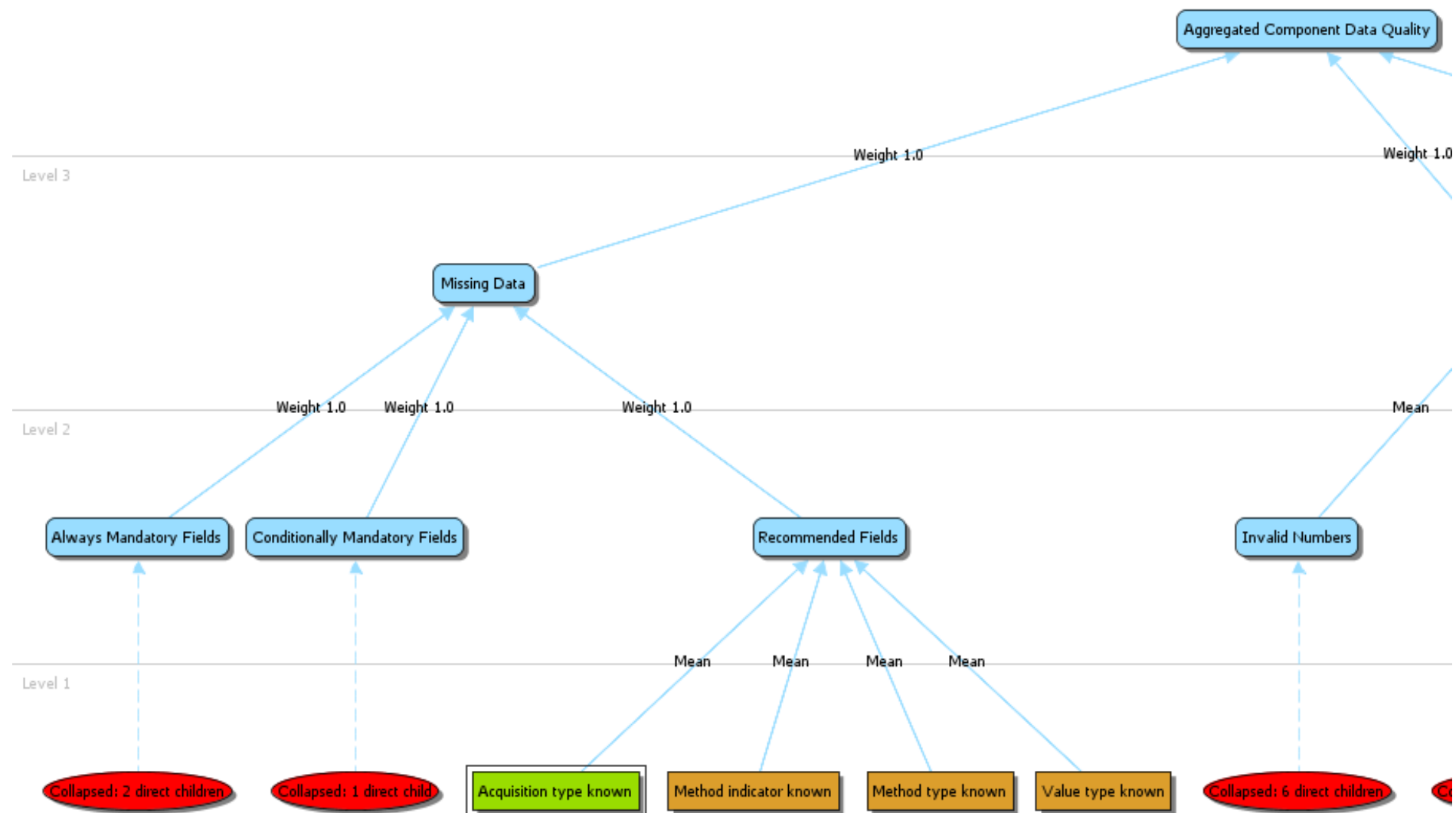
Assessment SQL

```
select {0},{2},
       case when
           aggrfoodcompmean is null or aggrfoodcompminimum is null
           or aggrfoodcompmean = '' or aggrfoodcompminimum = ''
           then null
       when cast(aggrfoodcompmean as numeric) < cast(aggrfoodcompminimum as numeric)
           then 0.0
       else 1.0 end as val
from {1}
where {3}
```

Variables {0} = <requirementId>, <assessmentId>; {1} = <entity> <alias>; {2} = <keycolumn>; {3} = <condition>; {4} = <alias>

156 rows Open New Delete

DQAF Data Quality Requirements (DQR) – Tree Editor Example



DQIF Data Quality Requirements (DQR)

```
public class AcquisitionTypeValidator extends AbstractComponentValidator {

    @Override
    public boolean hasError() {
        return false;
    }

    @Override
    public boolean hasSuggestion() {
        if (acquisitionType.getValue() instanceof Acquisitiontype) {
            Acquisitiontype t = (Acquisitiontype) acquisitionType.getValue();

            if (t.is(Acquisitiontype.CodeEnum.UNKNOWN)) {
                return true;
            } else {
                return false;
            }
        } else {
            return false;
        }
    }
}
```

Food Composition (FC) & Total Diet Study (TDS)

Food Composition data (FC data)

- Nutritionally important components of foods
- Values for energy and nutrients
- Food composition databases (FCDBs)

Total Diet Study data (TDS data)

- Similar to FC data but containing contaminant analysis values instead of nutrient values
- TDS process is different to the FC process



Nutrient (beneficial): Vitamin C, protein
Contaminant (harmful): Mercury (heavy metal)

Agenda

- Motivation
- How to ensure data quality in an information system?
- Data quality input feedback (DQIF) & Data quality analysis feedback (DQAF)
- Our approach with Hard- & Soft-Constraints
 - DQIF
 - DQAF
- Description of the evaluation
- Results and findings

Main Contributions