A Data Quality Test

Operationalize two data quality dimensions of ISO 25012: Efficiency and Precision

Arjan Vis
Supervisor university: Dr Davide Ceolin
Supervisor SIG: Prof. Dr Joost Visser
Second reader: Dr Jan Wielemaker
Structure (1/2)

• Context & Problem Statement

• Research Question

• Methodology

• Measurements

• Calibration
Structure (2/2)

• Results

• Conclusion & Contribution

• Limitations

• Future Work

• Q&A
Context & Problem Statement (1/3)

• Relevant performance issue of:
  – Operating Processes
  – Decision-making Activities
  – Interorganizational cooperation requirements

• Data Quality problems cost businesses annually $600 billion

• Poor Data Quality leads to:
  – Customer dissatisfaction
  – Increased costs
  – Lower employee job satisfaction
Context & Problem Statement (2/3)

• ISO-25012
  – SIG Data Quality Model

• No standardized way of implementing the standard

• No golden measurements for the different dimensions
  – ISO-25024
Context & Problem Statement (3/3)

• Efficiency:
  – The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.

• Precision:
  – The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.
Research Question

• Which measurements should be used to operationalize the data quality dimensions, specifically Efficiency and Precision, as defined by the ISO?

• Sub-question 1: What measurements are developed by SIG and others in order to operationalize these dimensions?

• Sub-question 2: In what way could these measurements be improved?

• Sub-question 3: Is it necessary to develop different measurements for different databases?
  – Sub-sub-question 3a: If so, which measurements are needed for what kind of databases?
Methodology

- Literature research -> all the sub-questions
  Papers on data quality measurements

- Calibration -> sub-question 2 and 3
  Use different kind of datasets, change (the value of) various factors and measure the data quality

- Validation -> sub-question 2 and 3
  Validate the measurements on database(s) where data quality is already known of
Measurements – Efficiency

• Empty and NULL Cells

• Empty Tables
  – At least 10 tables in database

• Type Checking
Measurements – Syntactic Precision

• Invalid dates
  – Invalid years, months and days

• Primary keys

• Modus of length of fractional part
  – Modus Plus
Implementation – JAVA (1/2)

- Count the empty tables
  - SELECT COUNT(*)

- Count empty / NULL cells
  - SELECT COUNT(*) WHERE column IS NULL or column != ''

- Column type
  - Int: SELECT MAX, MIN
  - String: SELECT MAX(LENGTH, MIN(LENGTH)
Implementation – JAVA (2/2)

• Primary Keys
  – SELECT column_key, COUNT(*)

• Dates
  – SELECT DATE, COUNT(*)

• Decimals
  – SELECT LENGTH, LENGTH(SUBSTRING_INDEX, COUNT(*))
Implementation – R

• Created the risk-profile

• Created the star rating

• Write the results to CSV-files
Results Calibration – Efficiency

Risk-Profile Efficiency

- CheckType
- EmptyCell
- EmptyTable

Arjan Vis - Information Sciences MSc
Results Calibration – Efficiency

Risk-Profile Efficiency

Combined measurement
Results Calibration – Syntactic Precision

Risk-Profile Syntactic Precision

- IncorrectDay
- IncorrectMonth
- IncorrectYear
- PrimaryKeys
- DecimalsModus
- DecimalsModusPlus

Arjan Vis - Information Sciences MSc
Validation – Efficiency

• Expert opinion necessary to make clear that is correct
Validation – Syntactic Precision
Conclusion & Contribution (1/6)

Sub-question 1: What measurements are developed by SIG and others in order to operationalize these dimensions?

- **Efficiency:**
  - Runtime for a program, measurement or business process (Sadiq (2013), Saha et al. (2014), Panahy et al. (2014), Aggarwal et al. (2004))
  - “…efficiency refers to the ability of the customers to get to the Web site, find their desired product and information associated with it, and check out with minimal effort.” (Zeithaml et al., 2002)
  - Correct usage of data types, NULLs and empty cells in the database (SIG)
Results & Contribution (2/6)

Sub-question 1: What measurements are developed by SIG and others in order to operationalize these dimensions?

- **Syntactic Precision:**
  - Precision = TP / (TP + FP) (Batini et al., 2006).
  - \( CV_j = 100 \times \sqrt{\sum_{i=1}^{R} ((X_{ij}-X_j)^2/(R-1))/X_j} \) (Campana, 2001) \( \equiv \sigma /\mu \)
  - Li et al. (1994) and Sebastian-Coleman (2012):
    - Character fields:
      - Ratio of numerical characters / total number of characters
      - Ratio of white space characters / total number of characters
      - Average, variance and coefficient of variance of used length of characters / total maximum length
    - Numeric fields:
      - Average, variance and coefficient of variance of the column
      - Grouping
  - Incorrect format of dates
Results & Contribution (3/6)
Sub-question 2: In what way could these measurements be improved?

• Not suitable for the dimensions:
  – Not focused on data from databases but data in general
  – General measurements that needs additional information
  – Very specific measurements that cannot be generalized

• Created own measurements based on existing measurements
Results & Contribution (4/6)

Sub-question 3: Is it necessary to develop different measurements for different databases?

No

- ISO 25012 / ISO 25024

Yes

- Portability
  - Compliance to Core SQL:2011 or ISO/IEC 9075:2011
Results & Contribution (5/6)

Sub-sub-question 3a: If so, which measurements are needed for what kind of databases?

- Look at characteristics of other DBMS
- Adjust measurements to those characteristics
Results & Contribution (6/6)

Research question: Which measurements should be used to operationalize the data quality dimensions as defined by the ISO?

• **Efficiency:**
  - Type checking
  - Empty cells
  - Empty tables

• **Syntactic Precision:**
  - Incorrect dates
    - Days, months, years
  - Primary keys
  - Fractionals
    - Modus
    - Modus Plus
Limitations

- ISO 25024
- Number of (open source) MySQL datasets
- One database for validation
Future Work

• Find / improve / create measurements for the other dimensions of ISO 25012

• Use ISO 25024 to adjust / add measurements

• Make measurements compatible with other (non-)relational databases
Q&A