Visualization of Multi-Level Data Quality Dimensions with Qualle

Sheny Ilescas Martinez¹, Lisa Ehrlinger¹,², Wolfram Wöß¹

¹ Johannes Kepler University Linz, Austria
² Software Competence Center Hagenberg GmbH, Austria
Data Quality Research at JKU and SCCH

• Johannes Kepler University (JKU) Linz
  ◦ Senior researcher in research group of a.Univ.-Prof. Wolfram Wöß
  ◦ DQ tool QuaIIe: http://dqm.faw.jku.at (Ehrlinger et al. 2018)
  ◦ DQ tool DQ-MeeRKat: https://github.com/lisehr/dq-meerkat
  ◦ Talks at MIT Chief Data Officer and Information Quality Symposium 2019 and 2020

• Software Competence Center Hagenberg GmbH (SCCH)
  ◦ Lead of research focus “Data Management and Data Quality”
  ◦ Research on DQ issues with industrial companies (e.g., KTM)
  ◦ DQ tool: A DaQL to Monitor Data Quality in Machine Learning Applications
    • International Conference on Database and Expert Systems Applications. Springer, Cham (Ehrlinger et al. 2019)
Aim of this Research

• **Data quality** (DQ) assessment is challenging but necessary to ensure that (business) decisions derived from data can be trusted

• Different DQ dimensions and metrics have been developed (cf. Batini & Scannapieco 2016) and the **DQ tool QuaIIe** facilitates their calculation

• **Humans need to understand** these DQ metrics to make educated decisions

• We present a visualization approach to enable **human-centered DQ assessment** across multiple dimensions and arbitrary complex data sources
  ◦ Understandable design in web-based graphical user interface (GUI) that extends Qualle
  ◦ Management of data quality rules
  ◦ Scalability → valid solution for complex integrated information systems
  ◦ Trigger new DQ metric calculations
Related Work

• Related DQ tools with visualization approaches
  ◦ In contrast to Qualle, other tools consider only tabular data
  ◦ **Profiler**: DQ tool by Kandel et al. (2016)
    ▪ Visual assistance and automatic suggestion of visualizations for identifying problematic data
  ◦ **MetricDoc**: DQ tool by Bors et al. (2019)
    ▪ [https://github.com/christianbors/OpenRefineQualityMetrics](https://github.com/christianbors/OpenRefineQualityMetrics)

• Related research inspiring our visualization approach
  ◦ Abedjan et al. (2017) use sunburst diagram to visualize functional dependencies
  ◦ Xie et al. (2006) recommend hue for transmitting DQ information in multivariate data
  ◦ Gratzl et al. (2013) present an interactive visualization technique for rankings
The Data Quality Tool Qualle
(Data Quality Assessment for Integrated Information Environments)

• Java-based tool to estimate the quality of integrated information systems

• Advantages: domain independent and unsupervised

• Performs quality measurements on different aggregation levels

• Qualle implements DQ metrics for dimensions on
  ◦ **Instance-level**: accuracy, completeness, timeliness, and minimality
  ◦ **Schema-level**: completeness, correctness, minimality, normalization, pertinence, and readability
Components of Qualle

Data Quality Calculators
- Accuracy / correctness
  - RefCorrectnessCalculator(data)
  - RatioAccuracyCalculator(data)
  - DSDCorrectnessCalculator(schema)
- Completeness
  - RatioCompletenessCalculator(data)
  - UniqueRatioCompletenessCalculator(data)
  - FilledCalculator(data)
  - DSDCompletenessCalculator(schema)
- Pertinence
  - RatioPertinenceCalculator(data)
  - RatioPertinenceCalculator(schema)
- Timeliness
  - AverageCurrencyCalculator(data)
  - AverageTimelinessCalculator(data)
- Minimality / Duplicity
  - RecordMinimalityCalculator(data)
  - SchemaMinimalityCalculator(schema)
- Readability
  - SchemaReadabilityCalculator(schema)
- Normalization
  - NormalFormCalculator(schema)

Data Source Connectors
- ConnectorMySQL
- ConnectorCSV
- ConnectorOntology
- ConnectorCassandra
- ConnectorAlphavantage

Reporters
- XMLTreeStructureDQReporter
- ConsoleReporter
- CDQM reporter (DQ monitoring)

Lisa Ehrlinger - Visualization of DQ Dimensions with Qualle
Design Approach for Multi-Level DQ Dimensions

• General view
  ◦ Determine general quality state of an integrated information systems
  ◦ Identify focal points, i.e., sources that require attention
  ◦ Provide information of the DQ dimension on demand with tooltips

• Detailed view
  ◦ Navigate through details about DQ calculations
  ◦ Summary statistics of the DQ calculations
  ◦ Attribute information (if available)
General View

(A) Tree view
(B) Sunburst diagram
(C) Filter panel
(D) Resource loading and configuration
Sunburst Diagram in General View

• Qualle stores schema information of integrated information systems and their quality information in form of a tree

• Communicates hierarchical structure of the data
  ° Slices of inner circles have hierarchical relationships to segments of the outer circles
  ° Leaves of the tree are extreme outer parts of the graph

• Suitable for large trees
Assigning Color (Hue) to Sunburst Diagram

- Color palette that indicates meaning of DQ calculates supports user in DQ assessment
- In Qualle, multiple DQ dimensions can be assessed with one or several DQ metrics

<table>
<thead>
<tr>
<th>Categorization function</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute average per dimension</td>
<td>Accuracy avg: 0.3, Readability avg: 0.8, Completeness avg: 0.75</td>
</tr>
<tr>
<td>Compute average of all dimensions</td>
<td>Overall quality avg: 0.62</td>
</tr>
<tr>
<td>Determine the score and color of the result</td>
<td>0.62 -&gt; [0.5 - 0.75] “Good”</td>
</tr>
</tbody>
</table>

\[
ratings_s = \frac{\sum_{i=1}^{n} dim_{si}.w_i}{n}, \\
\]
\[
dim_{si} = \frac{\sum_{j=1}^{m} r_j}{m}, \\
\]
\[
\text{category}_s = \begin{cases} 
\text{poor,} & \text{if } ratings_s < 0.25 \\
\text{fair,} & \text{if } 0.25 \geq ratings_s < 0.5 \\
\text{good,} & \text{if } 0.5 \geq ratings_s < 0.75 \\
\text{excellent,} & \text{if } ratings_s \geq 0.75 
\end{cases}
\]

- \( rating_s \) = quality rating of element \( s \)
- \( w_i \) = weight of dimension \( i \)
- \( dim_{si} \) = dimension average of element \( s \) and dimension \( i \)
- \( r_j \) = rating computed with metric \( j \)
Implementation Architecture of Qualle Visualization Component

• External resources
  ◦ Data sources
  ◦ Existing DQ reports from Qualle

• Graphical user interface (GUI)

• Servers
  ◦ PHP for website
  ◦ Tomcat server for Qualle communication

• Qualle core to calculate DQ ratings
Format of QuaIIe DQ Reports

- Analyzing DQ reports is complex
- Requires an intuitive visual representation

DQ report fragment

```xml
<Concept URI="example/student/student" label="student">
  <Quality>
    <Ratings>
      <Accuracy>
        <Ratio>0.3</Ratio>
      </Accuracy>
      <Readability>
        <Ratio>0.8</Ratio>
      </Readability>
      <Completeness>
        <UniqueRatio>0.6</UniqueRatio>
        <Ratio>0.6</Ratio>
        <Mean_UniqueRatio>0.8</Mean_UniqueRatio>
        <Filledness>1.0</Filledness>
      </Completeness>
    </Ratings>
  </Quality>
</Concept>
```
Implementation Model of Visualization Component

- **Backend**
  - Java servlets from Java server pages (JSP) web application
  - PHP parser based on EBNF grammar

- **Frontend**
  - GUI components
Outlook

• **User experience evaluation** to determine how easy and efficient the execution of DQ measurement tasks are perceived

• GUI performance evaluation

• Extend GUI with capabilities to **support the visualization of continuous DQ measurements** over time
References


The research reported in this paper has been funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG.