Visualization of Multi-Level Data Quality Dimensions with Qualle

Sheny Ilescas Martinez[1], Lisa Ehrlinger[1,2], Wolfram Wöß[1]
[1] Johannes Kepler University Linz, Austria
[2] Software Competence Center Hagenberg GmbH, Austria

JOHANNES KEPLER UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

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 survey: https://arxiv.org/abs/1907.08138 (Ehrlinger et al. 2019)
  - DQ tool Qualle: 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 Qualle** 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

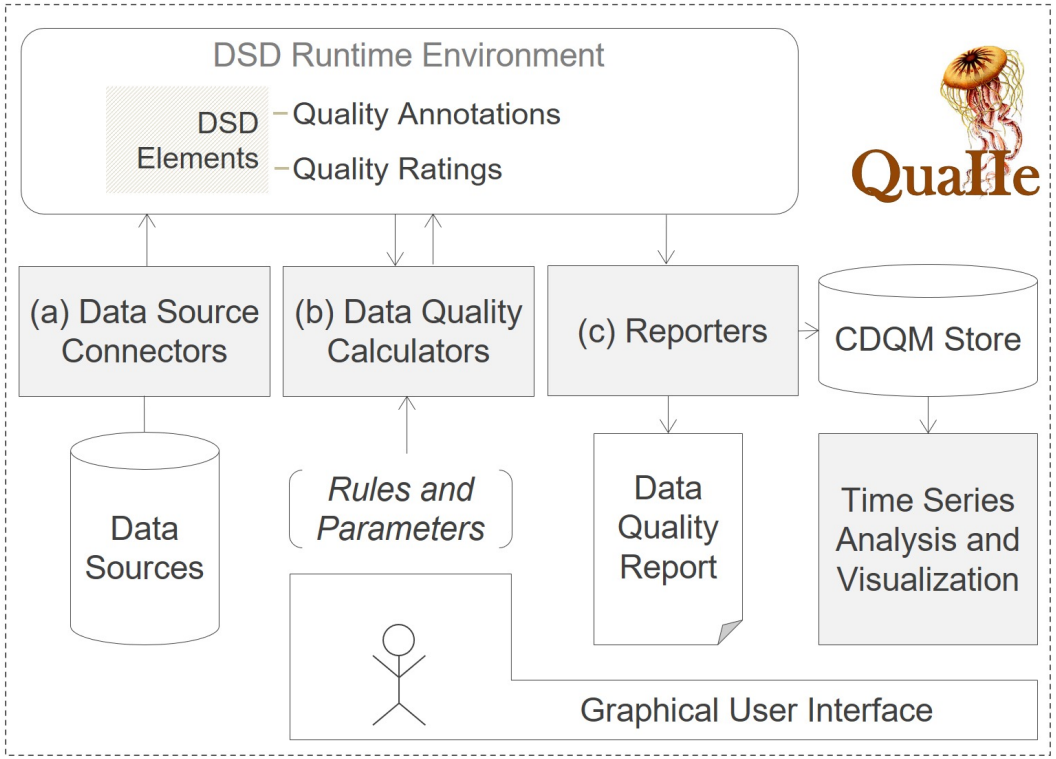
- Related DQ tools with visualization approaches
  - In contrast to Qualle, other tools consider only tabular data
  - **Profiler**: DQ tool by Kandel et al. (2016)
    - http://vis.stanford.edu/papers/profiler
    - Visual assistance and automatic suggestion of visualizations for identifying problematic data
  - **MetricDoc**: DQ tool by Bors et al. (2019)
    - 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

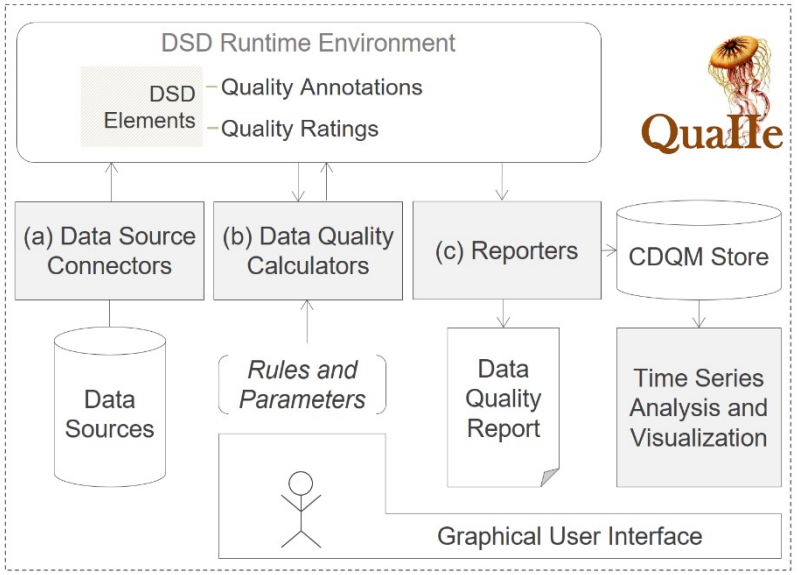
- Accuracy / correctness
  - `RefCorrectnessCalculator` (data)
  - `RatioAccuracyCalculator` (data)
  - `DSDCorrectnessCalculator` (schema)
- Completeness
  - `RatioCompletenessCalculator` (data)
  - `UniqueRatioCompletenessCalculator` (data)
  - `FilledCalculator` (data)
  - `DSDCompletenessCalculator` (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)

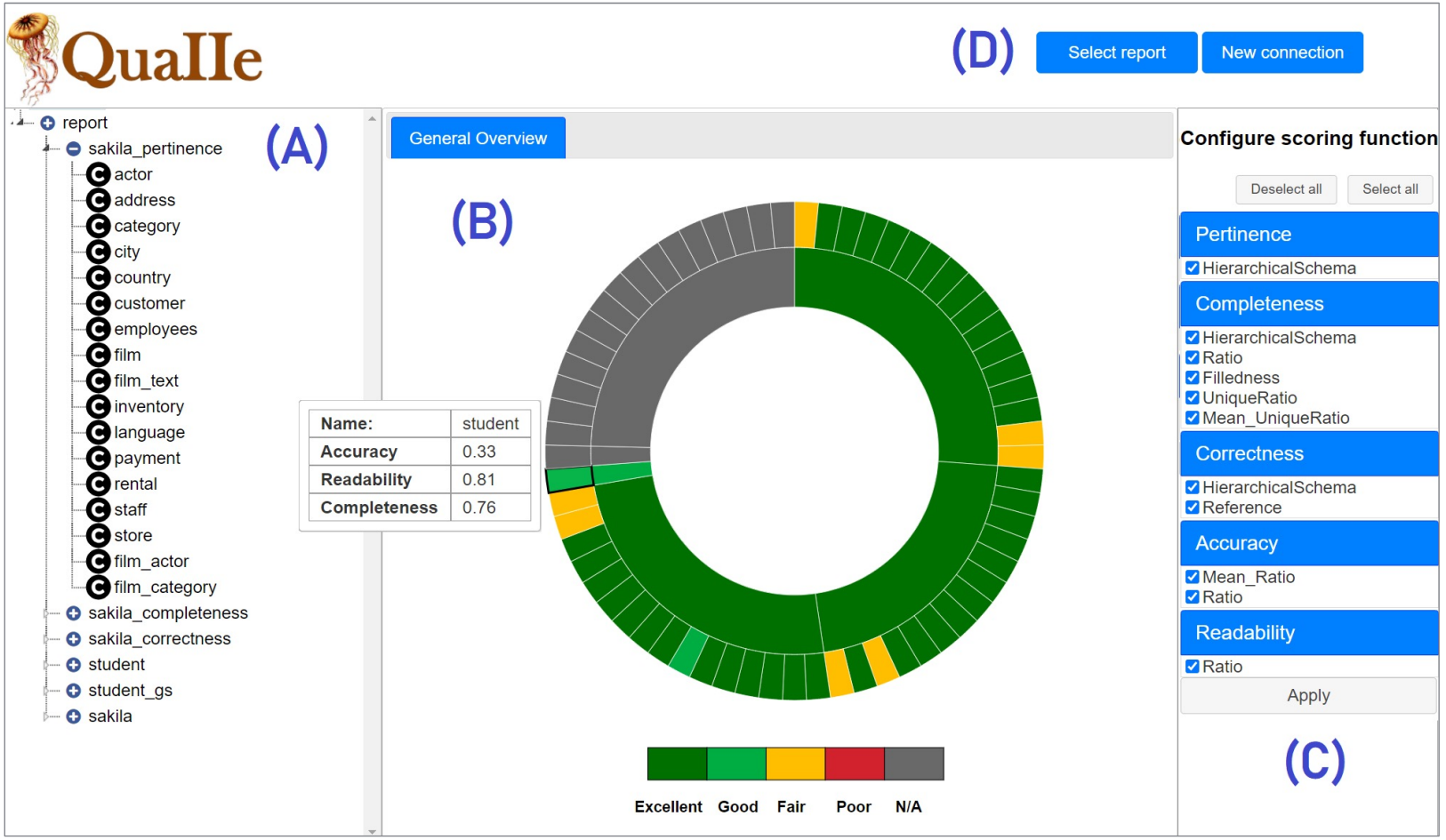


Design Approach for Multi-Level DQ Dimensions

- 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

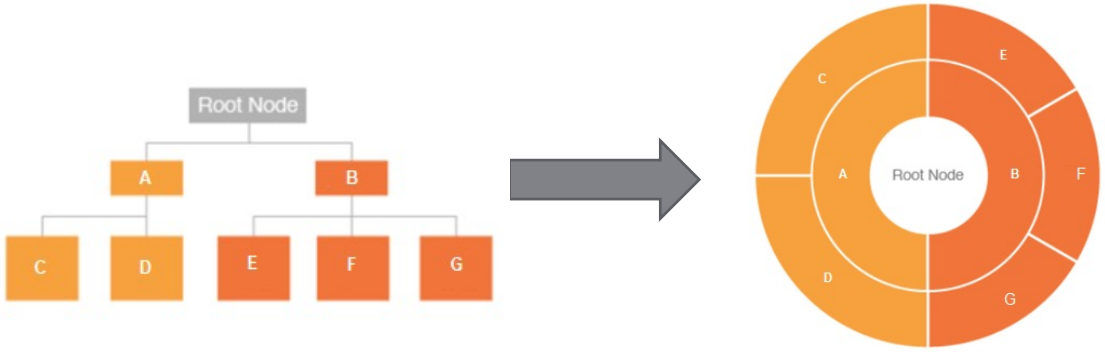
(B) Sunburst diagram

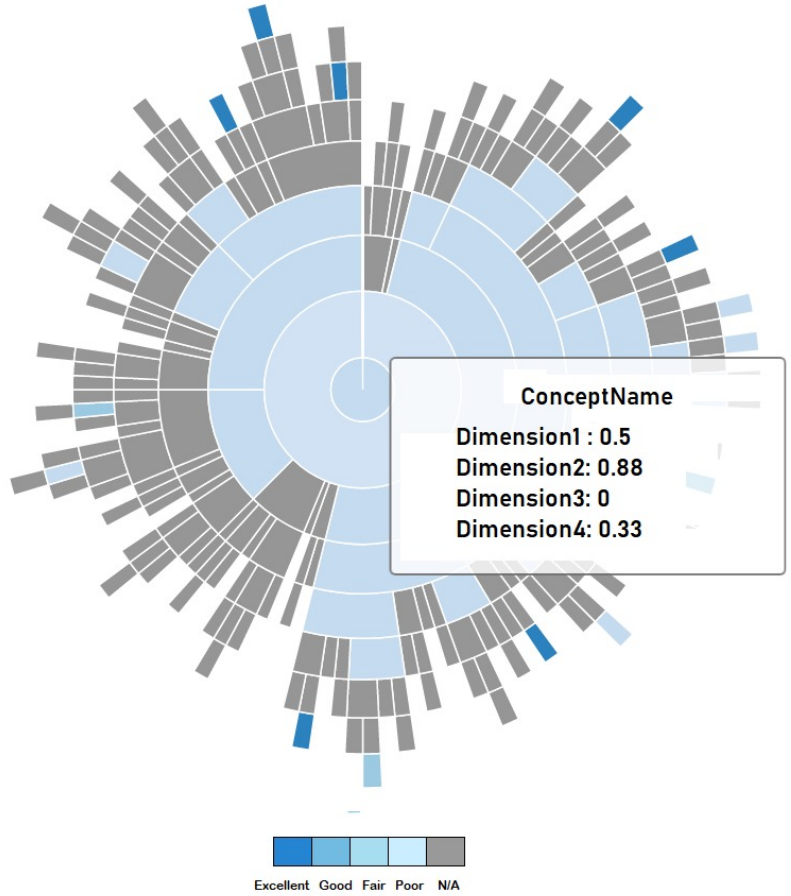
(C) Filter panel

(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

Categorization function	Example
Compute average per dimension	Accuracy avg: 0.3, Readability avg: 0.8, Completeness avg: 0.75
Compute average of all dimensions	Overall quality avg: 0.62
Determine the score and color of the result	0.62 -> [0.5 -0.75[“Good”

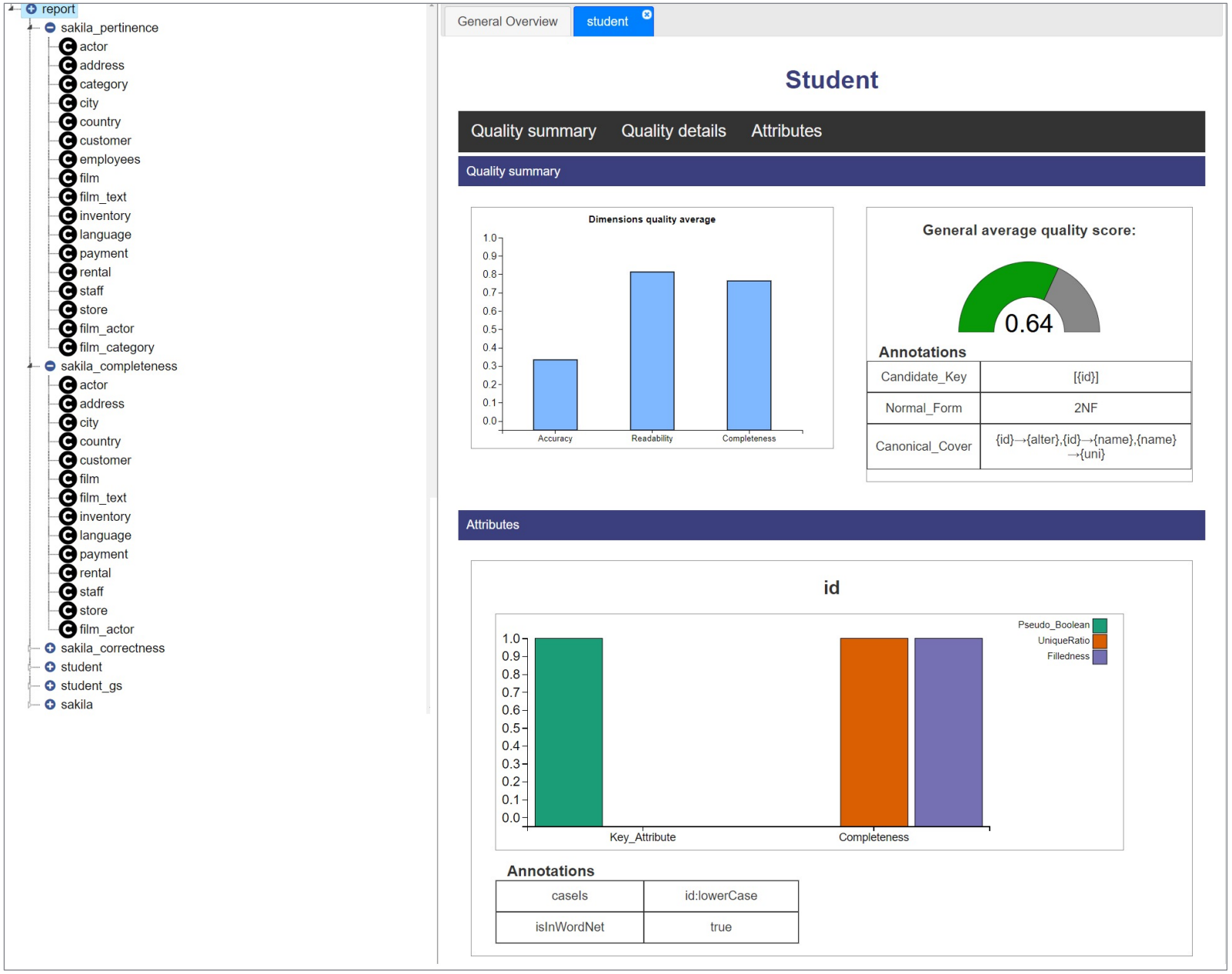
$$rating_s = \frac{\sum_{i=1}^{n} dim_{si}.w_i}{n},$$

$$dim_{si} = \frac{\sum_{j=1}^{m} r_j}{m},$$

$$category_s = \begin{cases} poor, & \text{if } rating_s < 0.25 \\ fair, & \text{if } 0.25 \geq rating_s < 0.5 \\ good, & \text{if } 0.5 \geq rating_s < 0.75 \\ excellent, & \text{if } rating_s \geq 0.75 \end{cases}$$

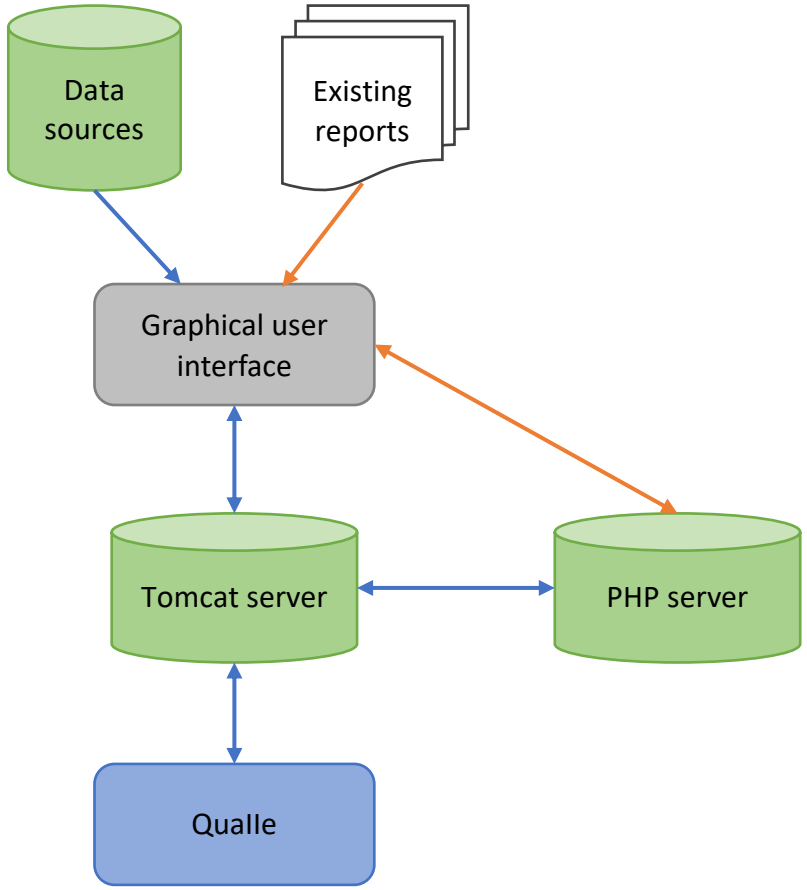
- $rating_s$ = quality rating of element *s*
- w_i = weight of dimension *i*
- dim_{si} = dimension average of element *s* and dimension *i*
- r_j = rating computed with metric *j*

Detailed View



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 Qualle DQ Reports

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

```
Report := Datasource {Datasource}
Datasource := [Quality] {Concept | RefAssociation}
Quality := Ratings | [Ratings] Annotations
Ratings := 0| 0.Digit{Digit} | 1
Digit := 0| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
Concept := [Quality] {Attribute}
RefAssociation := [Quality] {Attribute}
Attribute := [Quality]
```

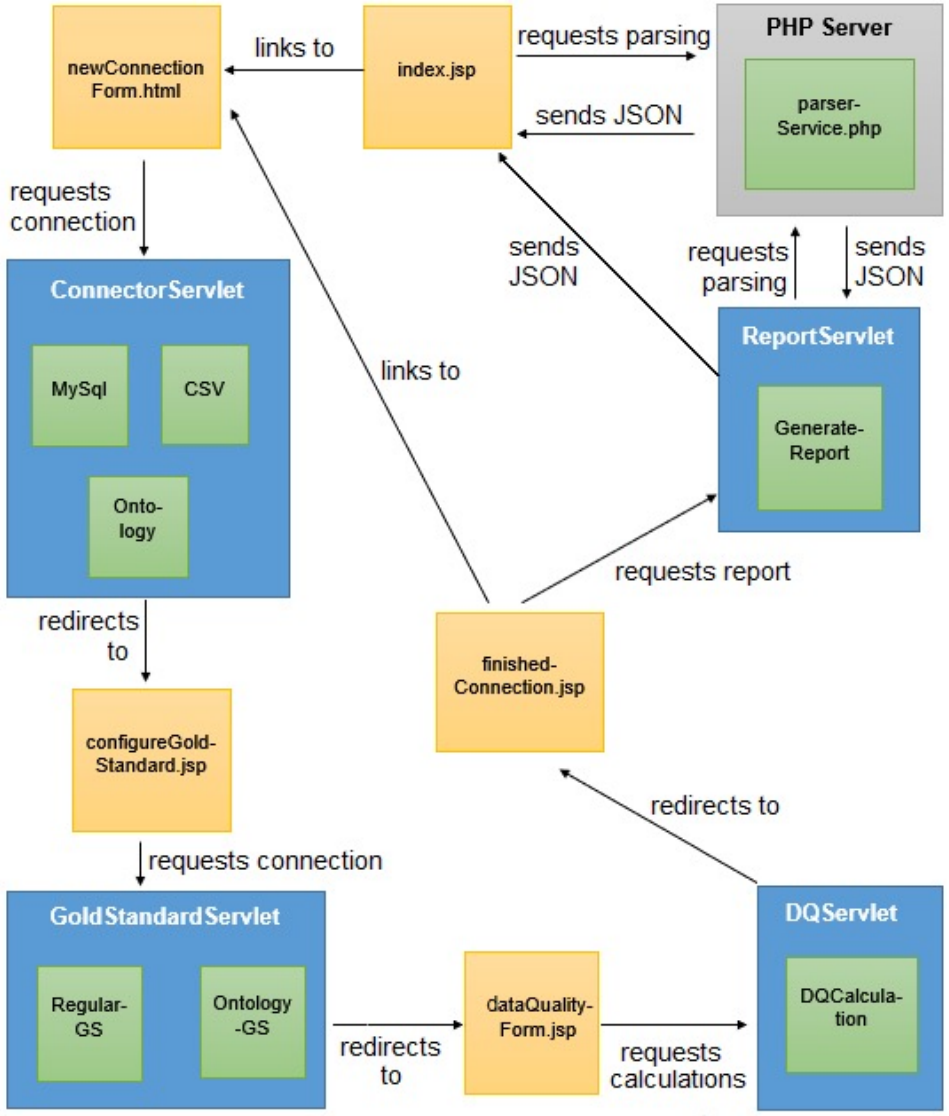
DQ report fragment

```
<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

- Z. Abedjan, L. Golab, and F. Naumann, “Data Profiling: A Tutorial,” May 2017, pp. 1747–1751.
- F. Aps, “Sunburst Diagram,” 2015. [Online]. Available: https://datavizproject.com/data-type/sunburst-diagram
- C. Batini and M. Scannapieco, Data and Information Quality: Concepts, Methodologies and Techniques. Springer International Publishing, 2016.
- C. Bors, T. Gschwandtner, S. Kriglstein, S. Miksch, and M. Pohl, “Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics,” Journal of Data and Information Quality, vol. 10, no. 1, pp. 3:1–3:26, May 2018. [Online]. Available: http://doi.acm.org/10.1145/3190578
- L. Ehrlinger, and W. Wöß, “Semi-Automatically Generated Hybrid Ontologies for Information Integration,” in Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems – SEMANTiCS2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15), vol. 1481, pp. 100–104, Nov. 2015.
- L. Ehrlinger, B. Werth, and W. Wöß. 2018. Automated Continuous Data Quality Measurement with QualIe. *International Journal on Advances in Software* 11, 3 & 4 (12 2018), 400–417.
- L. Ehrlinger, B. Werth, and W. Wöß. 2018. QualIe: A Data Quality Assessment Tool for Integrated Information Systems. In *Proceedings of the Tenth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2018)*. International Academy, Research and Industry Association, Nice, France, 21–31.
- S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, “LineUp: Visual Analysis of Multi-Attribute Rankings,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2277–2286, 2013.
- J. M. B. Josko and J. E. Ferreira, “Visualization Properties for Data Quality Visual Assessment: An Exploratory Case Study,” *Information Visualization*, vol. 16, no. 2, pp. 93–112, 2017. [Online]. Available: https://doi.org/10.1177/1473871616629516
- S. Kandel, R. Parikh, A. Paepcke, J. Hellerstein, and J. Heer, “Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment,” in *Advanced Visual Interfaces*, 2012. [Online]. Available: http://vis.stanford.edu/papers/profiler
- Z. Xie, S. Huang, M. O. Ward, and E. A. Rundensteiner, “Exploratory Visualization of Multivariate Data with Variable Quality,” in *2006 IEEE Symposium On Visual Analytics Science And Technology*, October 2006, pp. 183–190.

DI Lisa Ehrlinger
Senior Researcher, Johannes Kepler University Linz
lisa.ehrlinger@jku.at
Senior Researcher Data Science, Software Competence Center Hagenberg
lisa.ehrlinger@scch.at
+50 343 836
http://dqm.faw.jku.at

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.

Contact

JOHANNES KEPLER UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at