Institut für Parallele und Verteilte Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Datenqualitätsmetriken zur Unterstützung von Domänenexperten bei interaktiven Analysen

Jannis Rapp

Studiengang:	Data Science
Prüfer/in:	PD Dr. rer. nat. habil. Holger Schwarz
Betreuer/in:	Dipl.-Inf. Michael Behringer
Beginn am:	1. Januar 2020
Beendet am:	26. August 2020

Kurzfassung

Die in den letzten Jahren massiv angestiegenen Datenmengen führen zu zunehmenden Herausforderungen im Bereich der Datenanalyse. Automatische Methoden können bisher das für belastbare Analysen notwendige Domänenwissen nicht berücksichtigen, während gleichzeitig Domänenexperten immer häufiger eigenständige Analysen durchführen möchten. Der populäre Ansatz Self-Service-Business-Intelligence bietet hierfür jedoch zu wenig Freiheiten, weshalb häufig Data Mashup-Ansätze verwendet werden. Bei diesen stellen fehlerhafte Daten ein Problem dar, durch welches das Ergebnis von interaktiven Analysen wesentlich beeinflusst werden kann. Es ergibt sich die Herausforderung den Domänenexperten während der interaktiven Analyse zu unterstützen und so den Einfluss fehlerhafter Daten auf das Ergebnis der Analyse zu minimieren. In dieser Arbeit wird ein Konzept zur Unterstützung von Domänenexperten anhand einer in den Analyseprozess integrierten Datenqualitätsüberwachung entwickelt. Dieses Konzept definiert ein Prozessmodell für eine an den Bedürfnissen des Domänenexperten orientierte Überwachung der Datenqualität sowie Richtlinien für die Implementierung der enthaltenen Komponenten. Dieses wurde prototypisch in das an der Universität Stuttgart entwickelte Data Mashup-Werkzeug „FlexMash“ zur Modellierung von Datenflüssen implementiert und belegt die grundlegende Funktionalität des vorgestellten Ansatzes.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Motivation

Die Menge der weltweit erzeugten Daten verdoppelt sich alle zwei Jahre[1] und es wird geschätzt, dass der globale Datenbestand bereits im Jahr 2025 auf 175 Zettabytes angewachsen sein wird [RGR18]. Während die Datenmenge exponentiell ansteigt, bleibt die Geschwindigkeit, mit welcher diese von Menschen wahrgenommen werden können, allerdings nahezu konstant [MR10]. Automatisierte Ansätze können dieses Problem nicht lösen, da sie nicht in der Lage sind, Domänenwissen, Intuition und Entscheidungen in die Analyse zu integrieren [KMSZ09]. Die Integration von Domänenwissen in die Analyse ist jedoch entscheidend für die Belastbarkeit der Analyseergebnisse [BHM17].

Ein populärer Lösungsansatz für dieses Problem stellt die „Self-Service Business Intelligence“ (SSBI) dar [IW11]. Der SSBI Ansatz ermöglicht es Domänenexperten eigenständig personalisierte interaktive Analysen vorzunehmen [IW11]. Der Anwender ist unabhängig von der IT-Abteilung und kann deshalb die Erstellung des Ergebnisses um (bis zu) mehrere Monate beschleunigen [Eck09]. Die konventionellen SSBI Lösungen sind jedoch nicht flexibel genug und limitieren den Anwender während des Analyseprozesses auf vordefinierte Anwendungsfälle [BHM17; Hir15]. Um diese Limitierung zu umgehen, werden häufig Data Mashup-Werkzeuge verwendet [BHM17; Hir15].

Data Mashup-Werkzeuge gewähren dem Anwender mehr Freiheiten im Umgang mit Daten während der interaktiven Analyse. Daraus resultiert jedoch auch eine erhöhte Gefahr, während des Analyseprozesses fehlerhafte Daten zu verwenden. Durch die Verwendung fehlerhafter Daten während des Analyseprozesses sinkt die Belastbarkeit der Analyseergebnisse. Dies ist für Anwendergruppen mit wenig oder nur beschränkter Erfahrung im Umgang mit fehlerhaften Daten besonders problematisch.

In dieser Arbeit wird ein Ansatz zur Lösung dieser Problematik für die Anwendergruppe der Domänenexperten entwickelt. Als Domänenexperten werden in diesem Zusammenhang Anwender klassifiziert, welche über eine Expertise in dem Bereich, aus welchem die Daten erhoben wurden, verfügen. Im Allgemeinen können bei der Anwendergruppe des Domänenexperten daher keine tiefgreifenden technischen Kenntnisse vorausgesetzt werden. Für die Implementierung einer Datenqualitätsüberwachung in den interaktiven Analyseprozess eines Domänenexperten ergeben sich daher besondere Anforderungen. Diese werden im Folgenden anhand eines exemplarischen Anwendungsfalles veranschaulicht:

1.1 Exemplarischer Anwendungsfall

Für die Veranschaulichung der Implementierung der Datenqualität in den Analyseprozess des Domänenexperten wird ein Beispielszenario verwendet.

[1] EMC: The Digital Universe of Opportunities, https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Wartungsarbeiter	Datum	Maschine	Wartungsgrund	Fehlermeldung	Zustand
24	2013-12-21 09:24:00	42	warnung	F26	außer betrieb
5	2014-01-15 13:30:00	3	routine	null	funktionsfähig
17	2014-01-19 09:00:00	17	routine	null	funktionsfähig
13	2014-1-19 15:00:00	52	warnung	E02	defekt
2	2014-01-30 14:45:00	45	routine	null	funktionsfähig
17	2014-13-04 8:00:00	17	warnung	F19	außer betrieb

Tabelle 1.1: Beispiel für Wartungsdaten einer Firma

Domänenexperte für das Szenario ist ein Maschinenbauer. Dieser ist für den Betrieb von Maschinen in mehreren Fabriken verantwortlich. Der Maschinenbauer kennt sich bestens mit den Maschinen aus, verfügt jedoch über keinerlei Expertise im IT-Bereich. Er soll die Leistungsdaten dieser Maschinen überwachen. Für diese Aufgabe bekommt er von mehreren IT-Experten die Daten verschiedener Fabriken zur Verfügung gestellt. Die Daten setzen sich dabei aus maschinenbezogenen Daten (beispielsweise Luftdruck, Öltemperatur, RPM, Wartungsdaten...) sowie produktionsbezogenen Daten (beispielsweise Produkt, produzierte Einheiten, Produktionswarteschlange ...) zusammen. Ein Beispiel für mögliche Wartungsdaten stellt Tabelle 1.1 dar. Den Maschinenbauer interessieren bei der Auswertung der Daten mehrere Faktoren:

1. Zustand der Maschinen:

   Sind die Maschinen in fehlerfreiem Zustand oder muss eine Maschine gewartet und gegebenenfalls außer Betrieb genommen werden?

2. Auslastung der Produktionskapazität:

   Reichen die vorhandenen Maschinen aus? Muss die Produktionskapazität eventuell erhöht werden?

3. Gewinnung von Informationen:

   Geht Maschine XY bei der Fertigung von Produkt Z besonders häufig kaputt?

   Lassen sich (unbekannte) Anzeichen auf ein baldiges Versagen einer Maschine ermitteln?

   Sind sonstige nützliche Zusammenhänge in den Daten zu finden?

Die Implementierung eines Feedbacks zur Datenqualität in den vom Maschinenbauer modellierten Analyseprozess soll den Maschinenbauer während der eigenständigen Analyse unterstützen. Ziel ist es daher, den Maschinenbauer während dem Analyseprozess über Qualitätsprobleme der Daten

aufzuklären und die Korrektur dieser zu ermöglichen. Um dieses Ziel bestmöglich zu erfüllen, lassen sich mehrere Anforderungen an die Implementierung eines Feedbacks zur Datenqualität in den Analyseprozess des Maschinenbauers identifizieren:

A1 Integration in den vollständigen Datenanalyse-Prozess:

Der Maschinenbauer muss für ein erfolgreiches Ergebnis seiner Analyse auf mehrere Schritte des Datenanalyse-Prozesses (beispielsweise um die Daten vor der Auswertung zu bereinigen) zurückgreifen. Die Datenqualitätsüberwachung muss deshalb so in das System integriert sein, damit der Domänenexperten während des kompletten Datenanalyse-Prozesses unterstützt wird und die Auswirkungen der von ihm angewandten Operationen in dem angezeigten Feedback über die Datenqualität reflektiert wird. Bei Implementierungen, die nur einen Teil des Datenanalyse-Prozesses unterstützen, kann der Maschinenbauer genaue Problemstellen und die Auswirkung von angewandten Operationen schlechter identifizieren.

Diese Anforderung ist vor allem im Kontext seiner fehlenden Erfahrung mit der Analyse von Daten relevant, aufgrund welcher die Entstehung weiterer Datenqualitätsprobleme im Verlauf des Analyseprozesses (etwa durch die inkorrekte Anwendung von Operationen) wahrscheinlich ist.

A2 Feedback auf verschiedenen Granularitätsebenen:

An verschiedenen Stellen des Analyseprozesses interagiert der Maschinenbauer auf verschiedenen Abstraktionsebenen mit den Daten. Der Maschinenbauer interagiert beispielsweise bei der Modellierung seines Analyseprozesses mit den Daten auf einer hohen Abstraktionsebene (etwa mit Daten und Operationen als UI Elemente, welche per „Drag and Drop“ verbunden werden können) und bei der Transformation der Daten einer Datenquelle in tabellarischer Form auf einer niedrigen Abstraktionsebene.

Für die erfolgreiche Unterstützung des Maschinenbauers darf das Feedback über die Datenqualität daher nicht nur in eine Abstraktionsebene der Daten integriert werden. Für die Integration der Datenqualität auf den verschiedenen Abstraktionsebenen muss die Auflösung, mit welcher die Datenqualität angezeigt wird, abhängig von der Abstraktion mit welcher die Daten dargestellt werden angepasst, werden können.

Ein aggregiertes Feedback über die Datenqualität hilft dem Maschinenbauer beispielsweise bei der Modellierung des Analyseprozesses auf Probleme aufmerksam zu werden. Es ist jedoch nicht in der Lage, den Maschinenbauer auf die genauen Problemstellen der Daten in tabellarischer Form hinzuweisen.

Es ist daher essenziell, ein Feedback über die Datenqualität auf verschiedenen Granularitätsebenen zu implementieren.

A3 Beteiligung mehrerer Nutzergruppen:

Da der IT-Experte für das Management der Daten der Firma verantwortlich ist, kennt er sich am besten mit den Eigenschaften der Daten aus. Der Maschinenbauer hingegen kennt sich nicht mit den Daten, sondern mit der Domäne, aus der die Daten erhoben wurden, aus. Um ein optimales Feedback zur Datenqualität geben zu können, muss daher das Wissen beider Parteien bei der Ermittlung der Datenqualität kombiniert werden. Für den IT-Experten muss deshalb die Möglichkeit bestehen, sich mit seinem Wissen über die Daten am Analyseprozess des Maschinenbauers zu beteiligen.

A4 Verständliche Darstellung und Bedienung:

Da für den Maschinenbauer keine Expertise im IT-Bereich vorausgesetzt werden kann, müssen alle Elemente die zur Auswertung, Darstellung oder Veränderung der Datenqualität verwendet werden, einfach verständlich und bedienbar sein [BHM18]. Zu diesen Elementen zählen beispielsweise Dialoge zur Angabe von Parametern, Dialoge zur Visualisierung von Eigenschaften der Datenqualität und Operationen zur Transformation der Daten, welche die Datenqualität beeinflussen. Nur so kann sichergestellt werden, dass der Maschinenbauer ohne technische Expertise alle bereitgestellten Elemente verstehen und benutzen kann.

A5 Automatische Überprüfung der Datenqualität im Hintergrund:

Der Maschinenbauer verfügt aufgrund fehlender IT-Kenntnisse über keine Erfahrung mit der Datenqualitätsüberwachung, Datenqualitätsproblemen und dem Umgang mit Datenqualitätsproblemen. Eine selbständige, umfassende Überwachung der Datenqualität kann deshalb von ihm nicht erwartet werden. Das System muss die Überprüfung der Datenqualität deshalb automatisiert, ohne explizite Anweisung des Nutzers, sicherstellen. Nur so kann die Berücksichtigung der Datenqualität bei der Auswertung des Domänenexperten sichergestellt werden.

A6 Interaktion mit dem Domänenexperten:

Der Domänenexperte kennt sich am besten mit den Daten und dem Ziel seiner Auswertung aus [BHM18]. Dieses Wissen kann genutzt werden, um die Ermittlung der Datenqualität zu´verbessern und an den aktuellen Kontext anzupassen. Der Maschinenbauer kann beispielsweise ihm bekannte Zusammenhänge (etwa. Maschinentyp=„BO 570“ –> RPM<=3500) verwenden um Fehler in den Daten zu finden. Um dieses Wissen effektiv nutzen zu können, benötigt der Domänenexperte die Option mit dem System zu interagieren, um Parameter anzupassen und seine Expertise in den Datenqualitätsprozess einbringen zu können. Die Interaktion mit dem System zur Verbesserung der Datenqualität und die Anpassung an den Kontext der Auswertung wird daher vorausgesetzt.

A7 Korrektur von Datenqualitätsproblemen:

Der Maschinenbauer muss nicht nur auf Datenqualitätsprobleme aufmerksam gemacht werden. Für eine erfolgreiche Auswertung müssen diese auch behoben werden. Aufgrund seiner fehlenden Erfahrung im Umgang mit Daten kann die eigenständige Korrektur der identifizierten Fehler problematisch sein [LL19]. Um die Auswirkung von Datenqualitätsproblemen zu minimieren, darf das Datenqualitätsfeedback deshalb nicht nur Probleme der Daten beinhalten, sondern muss dem Nutzer auch aktiv Lösungen zu vorhanden Problemen vorschlagen. Die vorgeschlagenen Lösungen müssen vom System (auf Wunsch des Maschinenbauers) automatisiert ausführbar sein, um Fehler bei der Problembehebung weiter zu minimieren.

Diese Anforderungen werden als Grundlage für die Unterstützung des Domänenexperten während des Analyseprozesses verwendet.

1.2 Aufbau der Arbeit

Zunächst werden im Kapitel 2 grundlegende Begriffe, die für das Verständnis der folgenden Kapitel notwendig sind erklärt. Weiterhin werden in diesem Kapitel verwandte Forschungsarbeiten und kommerzielle Werkzeuge mit ähnlichen Zielsetzungen betrachtet. Kapitel 3 befasst sich mit der Unterstützung des Domänenexperten hinsichtlich der Datenqualität während der interaktiven Analyse. Im Zuge dessen wird ein Konzept für die Integration der Datenqualität in den Analyseprozess des Domänenexperten entwickelt. Darauf aufbauend wird in Kapitel 4 eine prototypische Implementierung des entwickelten Konzeptes präsentiert. Abschließend werden in Kapitel 5 die Ergebnisse dieser Arbeit zusammengefasst und weitere Möglichkeiten zur Weiterentwicklung aufgezeigt.

2 Grundlagen und verwandte Arbeiten

Dieses Kapitel erklärt grundlegende Begriffe und Konzepte, welche für das spätere Verständnis des vorgestellten Konzeptes vorausgesetzt werden. Des Weiteren werden in diesem Kapitel Verwandte Arbeiten vorgestellt und im Hinblick auf die Unterstützung des Domänenexperten analysiert.

2.1 Big Data

Die Datenmenge sowie die Geschwindigkeit mit welcher Daten produziert werden, steigt rapide an[1]. Der Begriff „*Big Data*“ bezieht sich auf die Eigenschaften, die sich für daraus für Daten ergeben und beim Umgang mit diesen berücksichtigt werden müsse [AA19]. Die drei populärsten Eigenschaften von „*Big Data*“ werden als die „3V's“ zusammengefasst [Lan01; AA19]. Diese beinhalten „Variety“ (Daten in unterschiedlichen Formaten z.B. strukturiert, semi-strukturiert und unstrukturiert), „Volume“ (bezogen auf die Menge der Daten) und „Velocity“ (die Geschwindigkeit, mit welcher neue Daten hinzukommen) [AA19]. Die „3V's“ können durch Hinzunahme der Eigenschaften „Veracity“ (Richtigkeit der Daten), „Validity“ (Tauglichkeit der Daten für einen bestimmten Kontext), „Volatility“ (Änderungsrate der Daten) und „ Value“ (Nutzen der Daten) auf „7V's “ erweitert werden [KUG14]. Wissenschaftliche Publikationen identifizieren immer weitere Eigenschaften von „*Big Data*“. Al-Mekhlal et al. [AA19] identifizieren in wissenschaftlichen Publikationen mittlerweile 15 Eigenschaften, die mit dem Begriff „*Big Data*“ verbunden werden können.

2.2 Explorative/Interaktive Datenanalyse

Explorative Datenanalyse bezieht sich auf die Analyse von Daten mit dem Ziel Zusammenhänge in den Daten (sogenannte „Muster“) aufzudecken [IPC15]. Im Gegensatz zur traditionellen Datenanalyse gibt es bei der explorativen Analyse keine feste Zielsetzung [IPC15]. Ziel der Analyse ist es stattdessen eine für den Analysten bisher unbekannte nützliche Information in den Daten zu finden [IPC15]. Die explorative Datenanalyse kann durch Integration interaktiver Elemente optimiert werden [IPC15]. Diese interaktiven Elemente können beispielsweise die Form von Visualisierungen oder „Exploration Interfaces“ annehmen, welche die Navigation der zugrundeliegenden Daten vereinfachen [IPC15].

[1] https://www.statista.com/statistics/871513/worldwide-data-created/

2.3 Mashups

Ein Mashup ist eine Anwendung, die sich aus mehreren wiederverwendbaren Komponenten zusammensetzt [DM14, Kapitel 1]. Durch die Kombination der Komponenten können oft neue Anwendungsgebiete erschlossen werden. Beispielsweise kann die Kombination verfügbarer Unterkünfte mit geografischen Daten das Finden einer Unterkunft vereinfachen [DM14]. Die Komponenten können in drei grundlegende Kategorien unterteilt werden [DM14, Kapitel 5]. Als „Data components“, welche Zugriff auf Daten liefern, „Logic components“, welche Zugriff auf Funktionalitäten, Unternehmenslogik oder Algorithmen liefern und als „User interface components“, welche ihr eigenes User Interface besitzen und sowohl Daten als auch eigene Anwendungslogik beinhalten können. Aufgrund ihres Aufbaus können Mashups flexibler eingesetzt werden, als herkömmliche Anwendungen und sind nicht auf spezifische Anwendungsfälle beschränkt [DM14, Kapitel 6].

2.4 Datenqualität

Der Einfluss der Datenqualität für Datenkonsumenten ist weitgehend anerkannt [CR19]. Zur Ermittlung der Datenqualität lassen sich in wissenschaftlichen Publikationen eine Vielzahl von Ansätzen finden [CR19]. Die Ansätze weichen dabei im verwendeten Auswertungsprozess, den Auswertungmethoden sowie den verwendeten Einflussfaktoren auf die Datenqualität voneinander ab [CR19]. In der wissenschaftlichen Literatur werden Eigenschaften der Daten, die einen Einfluss auf die Datenqualität haben als Datenqualitätsdimensionen bezeichnet [Zmu78]. Nachfolgend wird näher auf einzelne Dimensionen der Datenqualität eingegangen:

2.4.1 Datenqualitätsdimensionen

Die von Datenqualität Frameworks für die Beurteilung der Datenqualität genutzten Dimensionen weichen teilweise stark voneinander ab [CZ15]. Im Allgemeinen identifizieren wissenschaftliche Publikationen jedoch die Dimensionen Korrektheit, Vollständigkeit, Konsistenz und Aktualität als die wichtigsten Datenqualitätsdimensionen [ASWW18; JGDW18; SC02; BM11]. Datenqualitätsdimensionen können selbst wiederrum mehrdimensional sein und sich aus mehreren Datenqualitätselementen zusammensetzen [CZ15; BM09; Zmu78]. Nachfolgen werden die wichtigsten Datenqualitätsdimensionen kurz erklärt:

Korrektheit

Die Korrektheit umfasst die Abweichung der Attributwerte der Daten von den zugehörigen modellierten Entitäten der Real-Welt [Hin02; HKG12; Kai10; SMB05; WW96]. Scannapieco et al. [SMB05] unterscheiden weiterhin zwischen der syntaktischen und der semantischen Korrektheit. Die syntaktische Korrektheit erfasst, inwieweit die Werte eines Attributes von den korrekten Werten der Real-Welt abweichen [SMB05]. Die semantische Konsistenz erfasst bei der Korrektheit den Kontext der Benutzung eines Wertes [SMB05]. Werte können daher syntaktisch korrekt aber semantisch inkorrekt sein [SMB05]. **Beispiel:** Der Wert Autor=„J. R. R. Tolkien“ ist beispielsweise syntaktisch

korrekt aber im Zusammenhang mit dem Buchtitel=„Harry Potter“ semantisch inkorrekt. Der semantische Aspekt der Dimension Korrektheit überschneidet sich mit dem dem semantischen Aspekt der Dimension Konsistenz (Abschnitt 2.4.1).

Vollständigkeit

Die Vollständigkeit beschreibt den Grad, mit welchem die vorhandenen Daten ausreichen, um eine gegebene Aufgabe zu erfüllen [Kai10; BS06, Kapitel 2.2]. Für relationale Daten kann die Vollständigkeitsdimension anhand der vorhandenen NULL-Werte sowie der Annahme einer von zwei Vermutungen ausgewertet werden [BS06, Kapitel 2.2]. Die erste Vermutung nimmt die Vollständigkeit der vorhandenen Daten an und sieht alle Werte außerhalb der betrachteten Relation als falsch an („Closed World Assumption“) [BS06, Kapitel 2.2]. Wohingegen die zweite Vermutung keine Aussage über Werte außerhalb der betrachteten Relation erlaubt („Open World Assumption “) [BS06, Kapitel 2.2]. Mit der „Closed World Assumption“ wird also die Vollständigkeit der vorhandenen Relation bewertet und bei der „Open World Assumption“ werden auch Informationen außerhalb der betrachteten Relation berücksichtigt [SMB05; BS06, Kapitel 2.2]. Wären in der Relation „Deutsche Bundesländer“ beispielsweise 8 deutsche Bundesländer ohne NULL-Werte vertreten, so wäre die Vollständigkeit der Werte nach der „Closed World Assumption“ 8/8 und die nach der „Open World Assumption“ 8/16, da hier die nicht in der Relation vertretenen Bundesländer mit in die Vollständigkeit einfließen [SMB05] mit Bundesländern statt Studenten).

Für eine differenzierte Betrachtung der Vollständigkeit umfasst der Begriff des NULL-Wertes nicht nur Werte mit dem tatsächlichen Wert „NULL“, sondern bezieht sich auch auf andere Werte, welche die gleiche Semantik wie ein „NULL“-Wert (Wert nicht in verfügbar) besitzen (beispielsweise Platzhalter Werte wie Email=„xxx@yyy.com“ Beispiel aus [Kai10]) [Kai10; SMB05; HKK08]. Bei der Auswertung der Vollständigkeitsdimension kann der Grund für das Fehlen eines Wertes berücksichtigt werden [Kai10; SMB05]. Werte die Fehlen und auch in der Real-Welt nicht existieren, stellen daher keine Unvollständigkeit dar, da für einen in der Real-Welt nicht existierenden Wert auch keine Abbildung des Wertes in den Daten existieren kann [Kai10; SMB05].

Konsistenz

Häufig treten innerhalb eines Datensatzes Widersprüche auf. Die Konsistenz beschreibt, inwiefern Widersprüche in den Daten vorhanden sind [ASWW18; BS06]. Die Konsistenz von Daten kann mithilfe von sogenannter „Konsistenz-Regeln“ überprüft werden [Int17, Kapitel 13; BS06, Kapitel 2.4; Hin02]. Eine Konsistenz Regel beschreibt einen konstanten Zusammenhang von Datenelementen (beispielsweise eine Relation zwischen zwei Attributen, Postleitzahl=74385 –> Ort=Pleidelsheim). Wird eine Konsistenzregel verletzt, so liegt eine Inkonsistenz in den Daten vor [Int17, Kapitel 13; BS06, Kapitel 2.4]. Die Konsistenzdimension ist selbst mehrdimensional und kann in die drei Bereiche Integrität, semantische Konsistenz und representative Konsistenz unterteilt werden [BM09; LPFW06, Kapitel 4]. Im Nachfolgenden werden diese aufgeführt:

Integrität: Codd [Cod90, Kapitel 13.2] definiert für die Integrität von Daten eine Reihe von Integritätsregeln. Diese lassen sich in fünf verschiedene Typen unterteilen:

Semantische Konsistenz: Die semantische Konsistenz beschreibt das Ausmaß semantischer Inkonsistenzen in einem Datensatz [HBSA18]. Als eine semantische Inkonsistenz wird ein logischer Widerspruch zwischen mehreren Datenelementen bezeichnet [BM09; HBSA18; LPFW06]. Logische Widersprüche können mithilfe von Regeln über Beziehungen in den Daten ermittelt werden [HBSA18]. Diese Regeln haben die Form „statement1 –> statement2“ [HBSA18]. Ein Beispiel für eine solche Regel ist (Uni=Stuttgart und Studiengang=Data_Science) –> Fakultät=5. Hier kann aus dem Studiengang an der Universität Stuttgart ein Rückschluss auf die korrekte Fakultät gezogen werden. Ein Eintrag der Form (Uni=Stuttgart, Studiengang=Data_Science, Fakultät=9) ist deshalb semantisch inkonsistent.

Entitäts Regeln: Die Entitäts Regeln setzen die Eindeutigkeit aller Primärschlüssel voraus [LPW04; .] Des Weiteren fordern sie einen NOT NULL-Wert bei allen Bestandteilen von Primärschlüsseln [Cod90, Kapitel 8.2-8.6; LPW04; .]

Referentielle Integritätsregeln: Referentielle Integritätsregeln stellen sicher, dass zu jedem Fremdschlüssel der entsprechende Primärschlüssel zugeordnet werden kann [OG08; Cod90].

Domänen Integritätsregeln: Domänen Integritätsregeln erlauben es für jedes Attribut einen zulässigen Wertebereich (eine Domäne) zu definieren [Cod90, Kapitel 3.1-3.2]. Mehrere Attribute können sich eine Domäne teilen [Cod90].

Nutzer definierte Regeln: Die Regeln, die in diese Kategorie fallen, überschneiden sich mit denen der semantischen Konsistenz [HBSA18].

Repräsentative Konsistenz: Bezieht sich auf die Konsistenz im Datenformat [BM09; BS06]. Erhält der Maschinenbauer aus 1.1 beispielsweise Wartungsdaten, welche für das Attribut „Zeitstempel “ die Werte „Friday, 29 May 2015 11:30:00“ und „2018-07-12T015:50:00 “ beinhalten, ist dies eine Verletzung der repräsentativen Konsistenz.

Aktualität

Die Aktualität beschreibt, inwiefern Daten sich durch zeitliche Unterschiede von den Entitäten, welche sie modellieren, verändert haben[WW96; HKG12; SMB05; PLW02]. Die Aktualität hängt vom Kontext ab, in welchem die Daten verwendet werden [HK11; BS06, Kapitel 2; PLW02]. Dies liegt daran, dass Daten des selben Alters sich für verschiedene Anwendungsfälle unterschiedlich gut eignen [HK11]. Ein Beispiel kann anhand von über einem Jahr alten Finanzdaten eines Unternehmens gestellt werden. Ein Analyst will diese verwenden, um die Jahresbilanz des letzten Geschäftsjahres aufzustellen. Die Aktualität der Daten ist für ihn ausreichend, da sich die Finanzdaten des Unternehmens zwar mit der Zeit verändert haben, dies für ihn jedoch nicht relevant ist. Will jedoch ein anderer Analyst die aktuellen Quartalszahlen berechnen, so ist die Aktualität des Datensatzes ungenügend.

2.4.2 Quantifizierung

In Abschnitt 2.4.1 wurden für die Datenqualitätsdimensionen aufgezeigt. Um eine Aussage über die Datenqualität treffen zu können diese Dimensionen quantifiziert werden [HKKW07]. Zur Quantifizierung dieser Dimensionen können Datenqualitätsmetriken verwendet werden [ASWW18; HKKW07]. In wissenschaftlichen Veröffentlichungen wurden bereits verschiedene Metriken für eine derartige Quantifizierung einzelner Dimensionen definiert (z.B. in [HKKW07; Hin02; HKK09; ES07]). Für eine qualitative Quantifizierung der Datenqualität ist die Auswahl der Metrik(en) essentiell [Hün11; Los11; ES07; HKKW07; PLW02]. Es befassen sich deshalb mehrere Publikationen mit den Anforderungen an Datenqualitätsmetriken [Hün11; Los11; ES07; HKKW07; PLW02]. Heinrich et al. [HHK+18] analysieren die verschiedenen Anforderungen an Datenqualitäsmetriken und fassen diese zu fünf Anforderungen zusammen:

(A1): Existenz eines Minimums und Maximums: Die Werte der Metrik müssen durch ein Minimum und ein Maximum begrenzt sein. Es muss außerdem jeweils genau einen (d.h. mindestens und maximal einen) Zustand der Daten geben, in welchem der Wert der Metrik exakt das Minimum bzw. das Maximum annimmt.

(A2): Intervallskalierte Werte: Die Werte der Metrik müssen intervallskaliert (Quantitativ, Bestimmbarkeit der Gleichheit von Intervallen und Differenzen [Ste46]) sein. Das bedeutet, dass äquivalente Differenzen zwischen zwei Werten der Metrik dasselbe Ausmaß an Veränderung der Daten bedeuten [HKKW07]. Das höhere Skalenniveau der Ratioskala (Intervalskaliert + proportional [Ste46]) ist zur Erfüllung dieser Anforderung ebenfalls erlaubt.

Die Werte einer intervallskalierten Metrik, die zusätzlich die Anforderung A1 erfüllt, können mit der Formel $\frac{(DQ-m)}{(M-m)}$ (m=Minimum, M=Maximum, DQ=Metrik Wert) grundsätzlich als Bruchteil der maximalen Differenz interpretiert werden. Um die Interpretierbarkeit der Werte einer Metrik weiter zu verbessern, schlagen Heinrich et al. [HHK+18] eine Transformation von (nur) intervallskalierten Metriken zu ratioskalierten Metriken vor. Diese Transformation kann für alle Metriken, die A1+A2 erfüllen, vorgenommen werden. Werte rationalskalierter Metriken können mit von obiger Formel als Bruchteil der maximalen Datenqualität angegeben werden (da hier gilt m=0) und sind so besser interpretierbar.

(A3): Qualität der Konfigurations-Parameter und der durch die Metrik bestimmten Werte: Die von der Metrik verwendeten Parameter sowie die durch die Metrik ermittelten Werte müssen „qualitativ“ sein. Zur Beurteilung der „Qualität“ der Werte und Parameter werden die Qualitätkriterien „Objektivität“, „Zuverlässigkeit“ und „Korrektheit“ verwendet. Diese werden wie folgt definiert:

Objektivität: Der Grad, zu dem die Bestimmung der verwendeten Parameter und die Ermittlung von Werten durch die Metrik objektiv (frei von fremden Einflüssen) erfolgen.

Zuverlässigkeit: Der Grad, zu welchem die Werte (bestimmte Parameter und Ergebnisse der Metrik) reproduzierbar sind.

Korrektheit: Die Genauigkeit, mit welcher eine Metrik die von ihr modellierte Eigenschaft der Daten widerspiegelt. Bei der Bestimmung von Parametern bezieht sich die Korrektheit auf die Genauigkeit, mit welcher der Parameter die von ihm modellierte Eigenschaft beschreibt.

(A4): Aggregation der Werte: Die Metrik muss auf verschiedenen Granularitätsebenen der Daten angewendet werden können. Es muss einen Algorithmus für eine Aggregation der Werte der Metrik auf die nächsthöhere Granularitätsebene geben. Das Resultat der Aggregation muss konsistent mit der Berechnung der Metrik auf der höheren Granularitätsebene sein.

(A5): Wirtschaftliche Effizienz: Die Anpasssung der Metrik an einen spezifischen Kontext, sowie deren Anwendung muss im Hinblick auf Kosten und Nutzen effizient (erwartete Kosten < erwarteter Nutzen) sein.

2.5 Metriken

Zur Quantifizierung einzelner Dimensionen schlägt die wissenschaftliche Literatur oft unterschiedliche Metriken vor. Im Folgenden werden Metriken zur Quantifizierung einzelner Datenqualitätsdimensionen vorgestellt.

Aktualität

Probabilistischer Ansatz [HK11]: Quantifiziert die Aktualitätsdimension der Daten als Wahrscheinlichkeit, mit der Wert der Daten zum Auswertungszeitpunkt noch immer die Realität widerspiegelt. Die Metrik nimmt die exponentialverteilte Abnahme der Wertequalität an. Zur Modellierung der Wahrscheinlichkeit benötigt die Dimension das Alter der Werte seit Ermittlung als $age(\omega, A)$ (mit ω:=Attributwert und A:= Attribut) sowie die Veränderungsrate der Werte. Die Veränderungsrate eines Attributes über die Zeit wird von der Metrik als durchschnittliche prozentuale Veränderung pro Zeitschritt als $decline(A)$ modelliert. Der Faktor $decline(A) = 0.2$ bedeutet also beispielsweise, dass nach einem Zeitschritt 20% der Werte des Attributes A in den Daten von den der Real-Welt abweichen [HK11]. Zur Quantifizierung der Aktualität einzelner Werte eines Attributes wird von B. Heinrich und M. Klier [HK11] die Formel 2.1 verwendet.

$$Q_{Curr}(\omega, A) := exp(-decline(A) \cdot age(\omega, A)) \tag{2.1}$$

Zusätzlich erlaubt die Metrik eine Anpassung an den Kontext in dem sie angewendet wird durch Gewichtung der Attribute. Die Aggregation der Quantifizierungen auf die nächsthöhere Ebene ist durch den gewichteten Durchschnitt gegeben.

Zeitlich begrenzter Ansatz [BWPT98]: Der zeitlich begrenzte Ansatz nimmt an, dass Werte eines Attributes nach einem festen Zeitraum zwingend ungültig werden und modelliert den Verfall der Daten bis zu diesem Punkt. Die Aktualitätsdimension wird anhand des Alters der Daten von der Erhebung an ($currency$) und der Zeitspanne, welche die erhobenen Daten gültig bleiben, ($shelf\ life$) quantifiziert. Die Quantifizierung erfolgt mit:

$$Timeliness = max[(1 - currency/shelflife), 0]^s \tag{2.2}$$

Mit dem Parameter „s"“ wird die Sensibilität, mit welcher auf das „$currency/shelf\ life$“ Verhältnis reagiert wird, an den Kontext angepasst. Die „$shelf\ life$“ Variable muss an den Kontext angepasst werden (niedriger Wert für Attribute mit häufiger Änderung und hoher Wert für Attribute mit seltener Änderung).

Gewichtung nach „Update-Häufigkeit“ [Hin02] Quantifiziert die Aktualität eines Wertes an der „Update-Häufigkeit“ eines Attributes pro Zeiteinheit ($Upd(A)$) und der Zeit seit Erhebung des Wertes aus der Real-Welt ($age(\omega)$ mit ω:=Attributwert von A). Zur Quantifizierung wird die Formel 2.3 verwendet.

$$Q_{Zeit}(w, A) := \frac{1}{Upd(A) \cdot Age(\omega) + 1} \tag{2.3}$$

Es wird eine Aggregation der Metrikwerte von Attributwerten auf die Tupelebene mit einem nach Wichtigkeit einzelner Attribute gewichteten arithmetischen Mittel definiert. Außerdem eine Aggregation auf Relations- und Datenbankebene anhand des arithmetischen Mittels der jeweils niedrigeren Granularitätsebene.

Hybrider Ansatz [Eve05] Wählt abhängig von dem betrachteten Anwendungsfall, den zeitlich begrenzten oder den probabilistischen Ansatz aus. Dabei wird der probabilistische Ansatz für Anwendungsfälle ausgewählt, bei welchen die Definition einer festen zeitlichen Begrenzung nicht möglich ist. Dies ist beispielsweise bei Kundendaten der Fall, da hier davon auszugehen ist, dass sich mit der Zeit ein Teil der Daten in der Real-Welt ändert (etwa durch einen Umzug des Kunden) es kann jedoch kein fester Zeitpunkt angegeben werden, ab dem dieser Fall definitiv eingetreten ist. Die Zeitdauer kann daher nicht begrenzt werden.

Vollständigkeit

Verhältnis fehlender Werte [BKBJ14; BS06; HKK08; SMB05; Kai10]: Quantifiziert die Vollständigkeit der Daten als Verhältnis der vorhandenen Daten zu den fehlenden Daten (NULL-Werten). Für einzelne Attributswerte wird die Vollständigkeitsdimension mit einer 1 quantifiziert, wenn der Wert vorhanden ist, und mit einer 0, falls nicht (Q(Wert):=1 falls Wert != Null sonst 0)). Auf der Tupelebene entspricht das Verhältnis der Formel 2.4.

$$Q_{Voll}(Tupel) = \frac{\sum_i^{|Tupel|} Q(Tupel.Attribut_i)}{|Tupel|} \tag{2.4}$$

Heinrich et al. [HKK08] verwenden zusätzlich in 2.4 eine Gewichtung nach Attributen. Die Aggregation der Tupelebene auf die Relationsebene wird als Durchschnitt der Metrikwerte aller Tupel gebildet. Scannapieco et al. [SMB05] definiert die Vollständigkeit zusätzlich für einzelne Attribute als Verhältnis der nicht „NULL“ Werte eines Attributes zu der Gesamtzahl der Werte des Attributes.

Verhältnis der Tupel [BM11]: Quantifiziert die Vollständigkeit einer Relation als Verhältnis von Tupeln mit „NULL“ Werten zu Tupeln ohne „NULL “ Wert mit der Formel 2.5.

$$completeness = 1 - \frac{M_T}{N_K} \tag{2.5}$$

mit:
M_T := Anzahl der Tupel der Relation mit Nullwert
N_T := Anzahl der insgesamten Tupel der Relation
Die Existenz mehrerer „NULL“ Werte innerhalb eines Tupels wird bei dieser Quantifizierung vernachlässigt.

Die Quantifizierung der Vollständigkeitsdimension ist unabhängig von der gewählten Metrik auch jeweils von der Annahme der „Open World Assumption“ oder der „Closed World Assumption“ abhängig (siehe 2.4.1)

Korrektheit

Die Auswertung der Korrektheitsdimension ist schwierig, da für die Auswertung korrekte Referenzdaten benötigt werden [Los11, Kapitel 8]. Die vorgestellten Metriken setzen die Identifizierbarkeit inkorrekter Elemente in den Daten voraus.

Verhältniss korrekter zu inkorrekten Werten [SETN16; ASWW18; Jud15; Eve05]: Quantifizieren die Korrektheit als Verhältnis der korrekten Werte zu der Gesamtanzahl der Werte.

$$Accuracy = \frac{numOfCorrectValues}{totalValues} \tag{2.6}$$

Abstandsfunktion [CDFS11; Hin02; Eve05]: Berücksichtigt bei der Quantifizierung nicht nur die Korrektheit/Inkorrektheit der Werte, sondern auch den Grad zu dem ein Wert inkorrekt ist. Für die Bestimmung der Ähnlichkeit von Werten wird eine Abstandsfunktion benutzt. Als Abstandsfunktionen für Werte im String Format können beispielsweise die Levenshtein- oder die Hamming-Distanz verwendet werden. Zur Quantifizierung wird eine Formel 2.7 von B. Carlo et al. [CDFS11] benutzt.

$$Acc(t) = \frac{\sum_{|t|}^{i=1} acc(r_i, D(r_i))}{|t|} \tag{2.7}$$

$$\text{und } acc(r_i, D(r_i)) = \begin{cases} 1 & \text{if } r_i \in D(r_i) \\ 1 - NED(r_i, D(r_i)) & \text{otherwise} \end{cases} \tag{2.8}$$

Mit:

$Acc(t)$: = Quantifizierte Korrektheit
t : = Tupel
r_i : = ite Element aus t
D : = Domäne mit korrekten Werten
NED : = Abstandsfunktion (in diesem Fall Normalized Edit Distance)

Inkorrekte Werte, welche einen Rückschluss auf den korrekten Wert zulassen, wie einfache Typos (beispielsweise der Wert Land=„Detschland“), werden hier bei der Auswertung berücksichtigt. Der Ansatz der Quantifizierung der Korrektheit mithilfe einer Distanz Funktion eignet sich nur für den syntaktischen Aspekt der Korrektheitsdimension [BS06].

Konsistenz

Verhältnis konsistenter zu inkonsistenter Instanzen [LPFW06]: Quantifiziert die Konsistenz als Verhältnis der überprüften Instanzen zu den Instanzen, welche die Konsistenz verletzten.

$$Consistency = 1 - \frac{Number\ of\ instances\ violating\ a\ specific\ consistency\ type}{Number\ of\ instances\ checked} \tag{2.9}$$

Gewichtete Summe[AW14; HMHN07] Berücksichtigt bei der Auswertung der Konsistenzdimension auch die Konsequenzen der Erfüllung beziehungsweise der Nichterfüllung von Konsistenzregeln. Es wird hierfür eine Funktion w definiert welche für eine Regel r die Erfüllung/nicht Erfüllung/Unanwendbarkeit auf einen Wert abbildet. Die Konsistenz eines Tupels t wird nach 2.10 berechnet [AW14].

(2.10)

$$Consistency(t) = \sum_{r \in R} \begin{cases} w^{-}(r) & \text{if } t\ violates\ r \in D(r_i) \\ w^{+}(r) & \text{if } t\ fulfills\ r \\ w^{0}(r) & \text{if } r\ does\ not\ apply \end{cases}$$

Als Beispiel für eine mögliche Gewichtung der Regeln benutzen Alpar und Winkelstätter [AW14] die „confidence“ der jeweiligen Regel ($confidence() := support(X \Longrightarrow Y)/support(X) | support := prozentualer\ Anteil\ mit\ dem\ X\ zutrifft$). Hipp et al. [HMHN07] verwendet ausschließlich die „confidenece“ mit $w^{+} := -confidence^{\tau}$ und $w^{-} := confidence^{\tau}$. Bei Verwendung der „confidence “ zur Gewichtung wird sowohl von Alpar und Winkelstätter [AW14] als auch von Hipp et al. [HMHN07] ein Exponent τ benutzt, um geringere „confidence“ Werte überproportional schwächer zu gewichten.

CFDs maximale konfliktfreie Menge [WLG16] Stellt einen Ansatz zur Quantifizierung der Konsistenz für „Conditional Functional Dependencys“ vor. Die Quantifizierung basiert auf dem Verhältnis der minimalen Anzahl an Tupeln, welche aus den Daten D entfernt werden müssen, damit alle CFDs aus der Regelmenge (Σ) erfüllt werden, zu der Anzahl der insgesamten Tupel |D|. Für die Quantifizierung definiert Wang sogenannte „culprits“ ($C(D)$ mit $D - C(D) \models \Sigma$) als eine Teilmenge der Tupel, welche als Differenz der Tupel in D alle Regeln in Σ erfüllt. Des Weiteren definiert er $C_{min}(D)$ ($C_{min} := \forall C(D), |C_{min}(D)| \leq |C(D)|$) als kleinste entfernbare Menge. Die Konsistenz kann nun mit der Formel 2.11 (aus [WLG16]) quantifiziert werden.

(2.11)

$$consistency(D, \Sigma) = 1 - \frac{|C_{min}(D)|}{|D|}$$

2.6 Voraussetzungen für einen interaktiven Data Mining Prozesses

Behringer et al. [BHM18] stellen Rahmenbedingungen für interaktive Analyseprozesse, bei denen der Domänenexperte im Mittelpunkt steht. Es werden 5 Anforderungen an den Analyseprozess gestellt, um Domänenexperten eine erfolgreiche interaktive Analyse zu ermöglichen:

1. Put the User in Charge

Der Domänenexperte kennt sich am besten mit den Anforderungen und Erwartungen für seinen Anwendungsfall aus. Er sollte deshalb vollständige Kontrolle über jeden Schritt des Prozesses (von der Selektion der Daten bis zum Ergebnis) haben.

2. Explorative Character

   Durch die Kombination von Daten- und Verarbeitung können sich die Eigenschaften der Daten während eines Prozesses ändern. Für eine erfolgreiche Analyse und Verarbeitung der Daten benötigt der Domänenexperte ein weitreichendes Verständnis über die Daten. Er muss also die Option haben, diese zu „explorieren“.

3. Reduction of Complexity

   Die für den Kontext der Benutzung relevanten Abläufe des Systems müssen für Nutzer ohne Expertise im Bereich Informatik verständlich (der Nutzer versteht was das System macht) und benutzbar (die Anwendung/Benutzung eines Algorithmus darf keine komplexen Parameter benötigen) bleiben.

4. Balance of Techniques

   Die Balance zwischen automatischer Analyse und interaktiver Visualisierung muss so gewählt werden, dass die anderen Prinzipien weiterhin erfüllt werden. Idealerweise ist der Grad der Automatisierung nutzerabhängig.

5. Generic Approach

   Aufgrund verschiedener Anwendungsgebiete und Datenquellen ist ein generischer Ansatz wichtig. Nur so kann die Kompatibilität von Operationen zu Datenquellen sowie die Verbindung verschiedener Datenquellen gewährleistet werden.

2.7 FlexMash

FlexMash ist ein an der Universität Stuttgart entwickeltes Data Mashup Werkzeug (Abschnitt 2.3). Es erlaubt eine interaktive Verarbeitung und Analyse von Big Data 2.1. Zur interaktiven Analyse werden in FlexMash Datenflüsse modelliert. Für diesen Zweck stellt FlexMash eine Reihe von Daten-, Verarbeitungs- und Analyseknoten bereit, welche per „Drag and Drop“ aneinandergereiht und miteinander verbunden werden können. Der Fokus der Modelierung liegt auf einer einfachen intuitiven Benutzung, welche keine tiefergreifende Expertise in den Domänen der Informatik oder spezifischer der Datenverarbeitung und Analyse benötigt. Um dies zu ermöglichen, implementiert FlexMash die Anforderungen an die interaktive Datenanalyse aus dem Abschnitt 2.6.

2.8 Verwandte Arbeiten

Sowohl im wissenschaftlichen als auch im kommerziellen Bereich existieren eine Reihe von verwandten Arbeiten. Diese unterscheiden sich durch unterschiedliche Zielsetzungen. Die beiden Bereiche werden daher im Folgenden getrennt voneinander behandelt. Um zu beurteilen, ob der

Domänenexperte von der Integration der Datenqualität in verwandten Arbeiten adäquate Unterstützung während des Analyseprozesses erhält, werden die Anforderungen des Domänenexperten an die Integration der Datenqualität in den Analyseprozess (Abschnitt 1.1) verwendet.

2.8.1 Wissenschaftlicher Bereich

In wissenschaftlichen Publikationen wurden bereits eine Reihe von Ansätzen vorgestellt, welche die Expertise des Anwenders zur Auswertung der Datenqualität verwenden, oder den Nutzer mithilfe der Datenqualität bei der Datenanalyse unterstützen. Interessante Publikationen die diese Charakteristiken aufweisen umfassen HoloClean [RCIR17], ActiveClean [KFG+16], Wrangler [KPHH11], Xmdvtool[RWX+07], Falcon [HVS+16] und DQA[SBG+19]. Im Nachfolgenden wird ein kurzer Überblick über diese Veröffentlichungen gegeben:

HoloClean: HoloClean spezialisiert sich auf die probabilistische Korrektur von Datenqualitätsproblemen. Der Anwender interagiert mit HoloClean zur Detektion von Datenqualitätsproblemen durch die Spezifikation von Regeln, Abhängigkeiten und externen Referenzdaten. Bei der Reparatur von Problemen kann der Nutzer Reparaturmaßnahmen vor ihrer Anwendung überprüfen, um die Anwendung inkorrekter Korrekturmaßnahmen zu vermeiden [RCIR17].

ActiveClean: ActiveClean spezialisiert sich auf die iterative Bereinigung von Daten. Hierfür erstellt ActiveClean ein Modell für die Richtigkeit der zu bereinigenden Daten. Der Nutzer erhält dann eine Stichprobe der Daten, welche er durch die Transformation inkorrekter Werte verbessert. ActiveClean verwendet die vom Nutzer angewandten Operationen zur Verbesserung seines Modells für die Richtigkeit der Daten. Anschließend stellt es dem Nutzer eine neue Stichprobe zur Korrektur zur Verfügung. Das Modell wird so iterativ optimiert und abschließend zur Bereinigung des vollständigen Datensatzes verwendet [KFG+16].

Wrangler: Wrangler spezialisiert sich auf die Transformation von Daten. Der Nutzer wird bei der Transformation der Daten durch ein Attributspezifisches Feedback zur Datenqualität unterstützt. Die Auswirkungen von Angewendeten Operationen werden nicht nur in der Änderung der Daten, sondern auch in der Änderung der Datenqualität reflektiert [KPHH11].

Xmdvtool: Xmdvtool spezialisiert sich auf die visuelle Exploration von Daten. Die Datenqualität wird dazu verwendet erstellte Visualisierungen zu verbessern. Der Nutzer kann hierfür etwa mithilfe eines „Quality Brushes“ bei der Visualisierung nur Daten mit hoher Qualität berücksichtigen, oder Eigenschaften der Datenqualität auf grafische Eigenschaften (beispielsweise die Farbe) der Visualisierung abbilden („Visual Encoding“) [RWX+07].

Falcon: Falcon spezialisiert sich auf die Bereinigung der Daten. Zur Bereinigung der Daten interagiert der Nutzer auf den Daten in tabellarischer Form. Aus den Operationen, welche der Nutzer zur Korrektur von Datenqualitätsproblemen anwendet generiert Falcon verallgemeinerte Regeln (in Form von SQL Statements) zur Anwendung auf dem vollständigen Datensatz. Die generierten Regeln werden vor Anwendung vom Nutzer verifiziert, um Fehler zu vermeiden [HVS+16].

DQA: DQA wertet die Datenqualität auf Basis von Metadaten aus und generiert auf Basis der gefundenen Probleme Vorschläge zur Korrektur dieser. Der Nutzer stellt für diesen Prozess Metadaten zur Verfügung. Mithilfe der Metadaten wird die Datenqualität ausgewertet und ein

Feedback zur Datenqualität sowie eine Reihe von Korrekturvorschlägen erstellt. Zur Verbesserung der Datenqualität kann der Nutzer interaktiv aus verschiedenen Korrekturvorschlägen auswählen [SBG+19].

Die Tabelle 2.1 stellt die wissenschaftlichen Veröffentlichungen den Anforderungen des Domänenexperten aus Abschnitt 1.1 gegenüber. Nachfolgend wird näher auf die Konformität der wissenschaftlichen Veröffentlichungen zu den einzelnen Anforderungen eines Domänenexperten eingegangen:

A1 *Integration in den vollständigen Datenanalyse-Prozess*

Die Anforderung A1 wird von keinem der wissenschaftlichen Werkzeuge erfüllt. Dies ist vor allem auf die Spezialisierung auf eine bestimmte Aufgabe zurückzuführen. Diese sind daher nicht für die Überwachung der Datenqualität im vollständigen Analyseprozess geeignet.

A2 *Feedback auf verschiedenen Granularitätsebenen*

Die wissenschaftlichen Werkzeuge DQA und ActiveClean ermöglichen ein Feedback über die Datenqualität auf mehreren Granularitätsebenen durch die Verwendung von Übersichtsdialogen und der Darstellung konkreter Problemstellen in den Daten. Die Anforderung wird daher von diesen Werkzeugen erfüllt.

Wrangler erlaubt die Auswertung der Datenqualität für einzelne Attribute (Attributebene) und kann konkrete Problemstellen innerhalb der Attribute farbig hervorheben. Ein Feedback über die Datenqualität des vollständigen Datensatzes ist jedoch nicht implementiert. Die Anforderung wird von Wrangler deshalb großteils erfüllt.

Weitere Werkzeuge (Falcon, HoloClean) spezialisieren sich auf die Bereinigung und Korrektur der Daten und spezifizieren keine Aggregation des Feedbacks auf verschiedene Granularitätebenen. Die Anforderung wird von beiden Werkeugen daher nicht erfüllt.

A3 *Beteiligung mehrerer Nutzergruppen*

Keines der wissenschaftlichen Werkzeuge bindet mehrere Nutzergruppen in die Auswertung der Datenqualität ein. Es wird die Expertise eines einzelnen Nutzers zur Auswertung der Datenqualität verwendet. Die Anforderung wird deshalb von keinem Werkzeug erfüllt.

A4 *Verständliche Darstellung und Bedienung*

Falcon und Wrangler erfüllen die Anforderung, da die Werkzeuge für die Verwendung durch Nutzer ohne technische Expertise konzipiert sind. Die anderen Werkzeuge zielen auf Nutzer mit IT Kenntnissen ab und verwenden daher technisch komplexe Nutzerinteraktionen.

Das Werkzeug DQA, welches sich an Data Scientisten als Nutzergruppe richtet, und HoloClean, welches technisch komplexe Parameter wie beispielsweise „Matching Dependencys“ verwendet kann großteils ohne Expertise im IT-Bereich verwendet werden und erfüllen die Anforderung daher teilweise.

ActiveClean benötigt vom Nutzer unter anderem ein Modell zur Vorhersage der Korrektheit und einen Gradienten zur Optimierung des Modells. Da zur Angabe dieser Parameter eine technische Expertise notwendig ist, wird die Anforderung von ActiveClean nicht erfüllt.

Xmdvtool ist zur Erstellung komplexer Multivariater Visualisierungen konzipiert und eignet sich daher nicht für die Verwendung ohne statistische Vorkenntnisse.

A5 *Automatische Überprüfung der Datenqualität im Hintergrund*

Viele der wissenschaftlichen Werkzeuge (HoloClean, ActiveClean, Falcon, DQA) spezialisieren sich auf die Verbesserung oder Auswertung der Datenqualität. Ein solches Werkzeug muss vom Nutzer daher explizit mit der Intention, die Datenqualität sicherzustellen, gestartet werden. Die Überwachung der Datenqualität wird also nicht vom System automatisiert ausgeführt, da der Nutzer selbständig an die Verwendung der verschiedenen Werkzeuge denken muss. Die Anforderung wird von diesen Werkzeugen daher nicht erfüllt.

Wrangler erfüllt die Anforderung, da hier die Auswertung der Datenqualität automatisch vom System initialisiert wird. Der Benutzer muss nicht selbständig an die Verwendung der Funktion denken.

Xmdvtool integriert die Datenqualität automatisiert in die Visualisierungen des Benutzers. Auch hier ist die Anforderung erfüllt.

A6 *Interaktion mit dem Domänenexperten*

Fast alle wissenschaftliche Werkzeuge (außer Wrangler) verwenden die Expertise des Nutzers.

Wrangler besitzt jedoch keine Möglichkeiten für den Nutzer, die Auswertung der Datenqualität zu beeinflussen. Die Anforderung wird deshalb von Wrangler nicht erfüllt.

A7 *Korrektur von Datenqualitätsproblemen*

Fast alle wissenschaftliche Werkzeuge (Xmdvtool) bieten eine automatisierte Möglichkeit gefundene Probleme automatisiert zu korrigieren.

Die fehlende Option zur Korrektur von Datenqualitätsproblemen des Werkzeugs Xmdvtool kann auf die Spezialisierung des Werkzeugs auf die Visualisierung der Daten zurückgeführt werden.

Bei der Gegenüberstellung wird deutlich, dass keine Veröffentlichung die Anforderungen des Domänenexperten umfassend zufriedenstellt. Es zeigt sich vor allem der Bedarf nach einer „One Shot“ Lösung, welche ein interaktives Feedback zur Datenqualität in den kompletten Analyseprozess integriert. Sowie der Bedarf, mehrere Nutzergruppen an der Auswertung der Datenqualität zu beteiligen.

Tool	A1	A2	A3	A4	A5	A6	A7
HoloClean	✗	✗	✗	(✓)	✗*	✓	✓
ActiveClean	✗	✓	✗	✗	✗*	✓	(✓)
Wrangler	✗	(✓)	✗	✓	✓	✗	✓
Xmdvtool	✗	✗	✗	✗	✓	✓	✗
Falcon	✗	✗	✗	✓	✗*	✓	✓
DQA	✗	✓	✗	(✓)	✗*	✓	✓

* Werkzeug spezialisiert auf die Auswertung und/oder die Verbesserung

Tabelle 2.1: Konformität wissenschaftlicher Werkzeuge zu den Anforderungen des Domänenexperten aus Abschnitt 1.1

2.8.2 Kommerzieller Bereich

Außerhalb des wissenschaftlichen Bereichs existieren auch kommerzielle Werkzeuge welche Elemente der Datenqualität mit interaktiven Elementen verbinden. Aufgrund der Vielzahl an kommerziellen Produkten werden im Folgenden die Marktführer der beiden Bereiche Datenanalyse und Datenqualität nach Gartner[2] betrachtet. Diese werden Gartners „Magic Quadrants“ entnommen, welche vorhandene Anbieter in die vier Bereiche Marktführer, Herausforderer, Visionäre und Nischenanbieter unterteilt.

Im Bereich der Datenanalyse identifiziert Gartner die folgenden Marktführer (basierend auf dem „Gartner Magic Quadrant for Analytics and Business Intelligence“, Stand Januar 2020). Für diese werden jeweils die in der Kategorie „Analytics and Business Intelligence“ von Gartner vertretenen Produkte[3] betrachtet:

- Microsoft
- Tableau
- Qlik
- ThoughtSpot

Im Bereich der Datenqualität identifiziert Gartner die folgenden Marktführer (basierend auf dem „Gartner Magic Quadrant for Data Quality Tools“, Stand März 2019). Für diese werden jeweils die in der Kategorie „Analytics and Business Intelligence“ von Gartner vertretenen Produkte[4] betrachtet:

- Informatica
- IBM
- SAP
- SAS
- Syncsort
- Talend
- Oracle

Die kommerziellen Werkzeuge werden in Tabelle 2.2 ebenfalls den Anforderungen aus Abschnitt 1.1 gegenübergestellt. Nachfolgend werden die Ergebnisse der Gegenüberstellung für die einzelnen Anforderungen erklärt:

[2] https://www.gartner.com
[3] https://www.gartner.com/reviews/market/analytics-business-intelligence-platforms
[4] https://www.gartner.com/reviews/market/data-quality-tools

A1 *Integration in den vollständigen Datenanalyse-Prozess*

Die Integration der Datenqualitätsüberwachung in den vollständigen Analyseprozess wird von keinem der Anbieter im kommerziellen Bereich erfüllt. Dies lässt sich auf die Unterteilung des Analyseprozesses im kommerziellen Bereich zurückführen. Die Anbieter stellen für die einzelnen Bestandteile des Analyseprozesses separate spezialisierte Ansichten (oder Werkzeuge) zur Verfügung. Dies führt dazu, dass die Datenqualitätsüberwachung nicht in den vollständigen Analyseprozess integriert ist, sondern punktuell an Stellen des Prozesses angewandt wird.

A2 *Feedback auf verschiedenen Granularitätsebenen*

Die meisten kommerziellen Anbieter bieten eine differenzierte Betrachtung der Datenqualität auf mehreren Granularitätsebenen.

Eine Ausnahme Stellen die Anbieter Microsoft, Qlik und ThoughSpot dar. Diese spezialisieren sich im „Analytics and Business Intelligence“ Bereich auf die Analyse der Daten und bieten keine oder nur sehr eingeschränkte Datenqualitäts-Funktionalitäten.

A3 *Beteiligung mehrerer Nutzergruppen*

Bei der Anforderung A3 ist ein deutlicher Unterschied zu den Werkzeugen des wissenschaftlichen Bereichs (Abschnitt 2.8.1) erkennbar. Die kommerziellen Anbieter setzen fast vollständig auf eine Aufgabenverteilung zwischen verschiedenen Nutzergruppen oder empfehlen diese zumindest.

Bei der Auswertung dieser Eigenschaft muss angemerkt werden, dass die Expertise des Domänenexperten bei der Aufteilung der Nutzergruppen oftmals nicht für die Auswertung der Datenqualität beachtet wird.

A4 *Verständliche Darstellung und Bedienung*

Die Auswertung dieser Eigenschaft gestaltet sich schwierig. Bei vielen Anbietern sind jedoch technische Kenntnisse (etwa zur Definition von Modellen oder für die zur Erstellung von Scripts) zur Evaluation der Datenqualität nötig. Diese erfüllen die Anforderung daher nicht.

A5 *Automatische Überprüfung der Datenqualität im Hintergrund*

Bei allen Anbietern muss der Anwender aktiv an die Auswertung oder die Bereinigung der Datenqualität denken. Weil für den Domänenexperten ein System benötigt wird, welches den Prozess der Datenqualitätsüberwachung aktiv eigenständig anstößt, wird die Anforderung von keinem der Anbieter erfüllt.

A6 *Interaktion mit dem Domänenexperten*

Alle Anbieter können die Expertise von mindestens einer Anwendergruppe für die Auswertung der Datenqualität verwenden. Da viele der Anbieter jedoch keine Domänenexperten als Anwendergruppe spezifizieren gestaltet sich die Auswertung der Anforderung als schwierig. Die Anforderung gilt daher im Allgemeinen als erfüllt, da eine Interaktion mit den Anwendern stattfindet.

A7 *Korrektur von Datenqualitätsproblemen*

Fast alle Anbieter bieten dem Nutzer Möglichkeiten zur Korrektur gefundener Probleme.

Die Ausnahmen stellen erneut die Anbieter Microsoft, Qlik und ThoughSpot, da diese sich auf die Analyse der Daten spezialisieren und keine oder nur sehr eingeschränkte Datenqualitäts-Funktionalitäten bieten.

Anbieter	A1	A2	A3	A4	A5	A6	A7
Microsoft*	✗	✗	✓	✗	✗	✓	✗
Tableau	✗	✓	✓	✓	✗	✓	✓
Qlik	✗	✗	✗	✗	✗	✓	✗
ThoughtSpot	✗	✗	✓	✗	✗	✓	✗
Informatica	✗	✓	✓	✓	✗	✓	✓
IBM	✗	✓	✓	✓	✗	✓	✓
SAP	✗	✓	✓	✗	✗	✓	✓
SAS	✗	✓	✓	✗	✗	✓	✓
Syncsort	✗	✓	✓	✗	✗	✓	✓
Talend	✗	✓	✓	✓	✗	✓	✓
Oracle	✗	✓	✓	✓	✗	✓	✓

Tabelle 2.2: Konformität der Werkzeuge kommerzieller Anbieter zu den Anforderungen des Domänenexperten aus Abschnitt 1.1

Die betrachteten kommerziellen Anbieter erfüllen aufgrund ihres (durchschnittlich) größeren Produktumfangs mehr Anforderungen des Domänenexperten, als die wissenschaftlichen Werkzeuge. Es lassen sich jedoch auch mit den betrachteten kommerziellen Werkzeuge nicht alle Anforderungen des Domänenexperten zufriedenstellen. Auch hier zeigt sich der Bedarf nach einer „One Shot“-Lösung, welche die Datenqualität in jeden Schritt des Analyseprozesses integriert und automatisiert aktiv die Datenqualität bewertet und sicherstellt.

2.8.3 Prozessmodelle zur Auswertung der Datenqualität

Zusätzlich zu den wissenschaftlichen Werkzeugen (Abschnitt 2.8.1) und den kommerziellen Produkten (Abschnitt 2.8.2) stellen wissenschaftliche Publikationen eine Reihe unterschiedliche Methodiken zur Evaluation der Datenqualität dar [TKS+16; CZ15; SBG+19; LSKW02; CDFS11; VMH16; SETN16; MCR+16; AG09].

Im Hinblick auf die Anforderungen des Domänenexperten und der Herausforderungen von Big Data sind vor allem die Methodiken von Zai und Zhu [CZ15] sowie Taleb et al. [TKS+16] interessant.

Zai and Zhu [CZ15] stellen ein Prozessmodell für die Evaluation der Datenqualität von Big Data vor. Ziel des Prozessmodells ist die Sicherstellung einer für den Zweck der Analyse ausreichende Datenqualität [CZ15]. Das Prozessmodell eignet sich jedoch nicht zur Unterstützung des Domänenexperten, da es den Nutzer benötigt, um informierte Entscheidungen zur Auswertung der Datenqualität zu treffen (Anforderungen an die Datenqualität, relevante Dimensionen, Auswertung der Dimensionen,...). Diese kann der Domänenexperte nicht ohne technische Expertise treffen [CZ15]. Das

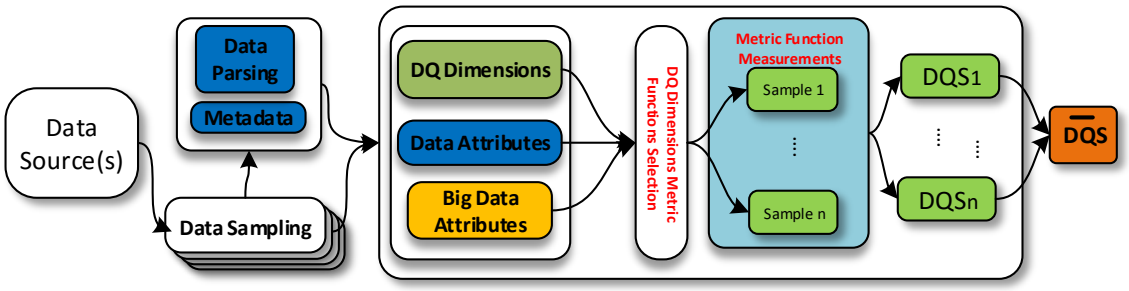

Abbildung 2.1: Prozess zur Stichproben basierten Auswertung der Datenqualität ([TKS+16])

Prozessmodell separiert zudem die Akquisition, Evaluation und Bereinigung der Daten zeitlich von der Analyse der Daten und kann somit nicht zur Unterstützung des Domänenexperten verwendet werden.

Talek et al. [TKS+16] evaluieren die Datenqualität basierend auf Stichproben der Daten, um die zur Auswertung benötigte Zeit und Rechenleistung zu reduzieren. Der hierfür verwendete Prozess (Abbildung 2.1) erstellt zuerst Stichproben aus den vorhandenen Datenquellen [TKS+16]. Aus diesen Stichproben und eventuell vorhandener Dateien, welche Metadaten beinhalten, werden durch eine automatische Analyse Metadaten generiert [TKS+16]. Diese Metadaten werden während der Auswertung als Parameter für die Auswertung von vordefinierten Datenqualitäts-Dimensionen (Talek et al. [TKS+16] verwenden Korrektheit, Konsistenz und Vollständigkeit) verwendet [TKS+16]. Bei den Parametern für die Auswertung von Datenqualitäts-Dimensionen unterscheiden Talek et al. [TKS+16] zwischen „normalen“ Eigenschaften der Daten (beispielsweise Datentyp und Format) und Big Data Eigenschaften (Größe der Daten, Anzahl der Datenquellen und Geschwindigkeit mit welcher neue Daten generiert werden). Diese Parameter werden für die Auswertung der Datenqualität auf den verschiedenen Stichproben verwendet und die Ergebnisse der einzelnen Dimensionen der Stichproben werden anschließend zu einem Endergebnis aggregiert.

Der Auswertungsprozess (Abbildung 2.1) ist auf die Stichproben-basierte Auswertung der Datenqualität fokussiert. Die Anforderungen des Domänenexperten (Abschnitt 1.1) an die Datenqualitätsüberwachung werden daher von dem Auswertungsprozess eigenständig nicht erfüllt, da für den Domänenexperten kritische Bestandteile im Prozess nicht spezifiziert sind (beispielsweise die Interaktion mit dem Domänenexperten, die Integration in den Analyseprozess, die Beteiligung mehrerer Nutzergruppen,...). Der Auswertungsprozess kann jedoch (in adaptierter Form) als Teil des Systems zur Unterstützung des Domänenexperten verwendet werden.

3 Konzept

Die Qualität der verwendeten Daten kann das Ergebnis datenbasierter Aufgaben maßgeblich beeinflussen [CR19]. Beim Analyseprozess des Domänenexperten muss diese daher berücksichtigt werden. Vorhandene Prozesse zur Überwachung der Datenqualität bieten dem Domänenexperten jedoch keine adäquate Unterstützung während des Analyseprozesses (Abschnitt 2.8). Es wird daher ein Konzept benötigt, welches auf die Anforderungen des Domänenexperten an die Integration der Datenqualität während des Analyseprozesses (Abschnitt 1.1) zugeschnitten ist.

Das Konzept umfasst ein Prozessmodell für die Integration einer Datenqualitätsüberwachung in den Analyseprozess des Domänenexperten. Es werden des Weiteren genaue Methodiken für die Interaktion mit und der Ermittlung der Datenqualität sowie die Integration dieser in den Analyseprozess des Domänenexperten definiert. Die Integration einer Datenqualitätsüberwachung nach diesem Konzept erfüllt die Anforderungen des Domänenexperten aus Abschnitt 1.1.

Im Folgenden wird zuerst das Prozessmodell zur Überwachung der Datenqualität vorgestellt, anschließend werden die einzelnen Komponenten des Prozessmodells genau spezifiziert.

3.1 Prozessmodell

Die Anforderungen des Domänenexperten an die Integration der Datenqualitätsüberwachung in den Analyseprozess sind wie in Abschnitt 1.1 aufgezeigt komplex. Damit die Datenqualitätsüberwachung diese Anforderungen erfüllen kann, müssen diese vom Prozessmodell erfüllt werden.

Für die Erfüllung der Anforderungen (Abschnitt 1.1) werden verschiedene Komponenten benötigt. Diese sind im Prozessmodell (Abbildung 3.1) mit K1-K5 gekennzeichnet und erfüllen jeweils spezifische Aufgaben während des Prozesses. Ein Überblick über die Verteilung der Zuständigkeiten wird nachfolgend gegeben:

Datenquellen-Repository (K1): Das Datenquellen-Repository verwaltet die Datenquellen, die während des Analyseprozesses verwendet werden können und die verfügbaren Metadaten über diese.

Datenqualitäts-Repository (K2): Das Datenqualitäts-Repository beinhaltet die Programmlogik, die zur Auswertung der überwachten Datenqualitäts-Dimensionen verwendet wird sowie eine Sammlung von Korrekturmaßnahmen für die während der Auswertung gefundenen Fehler.

Benutzeroberfläche (K3): Die Benutzeroberfläche ist für die Interaktion mit dem Domänenexperten zuständig. Die wichtigsten Aufgaben der Benutzeroberfläche sind die Erfassung von Metadaten vom Domänenexperten sowie die Übermittlung von Informationen über die Datenqualität an den Domänenexperten.

Datenqualitäts-Service (K4): Der Datenqualitäts-Service (K3) verwaltet den Informationsfluss zwischen den verschiedenen Komponenten des Prozessmodells um den korrekten Ablauf der Datenqualitätsüberwachung sicherzustellen.

Evaluationsumgebung (K5): Die Evaluationsumgebung erhält Informationen zur Auswertung der Datenqualität vom Datenqualitäts-Service. Die Informationen werden von der Evaluationsumgebung für die Auswertung der Datenqualität im Workflow des Domänenexperten verwendet. Dies resultiert in einem Feedback über die Datenqualität im Workflow des Domänenexperten.

3.1.1 Ablauf des Prozessmodells

Der Ablauf des Prozessmodells kann in zwei Phasen unterteilt werden, wobei sich die erste Phase mit den Interaktionen des IT-Experten befasst und die zweite mit denen des Domänenexperten. Die beiden Phasen werden im Folgenden näher beschrieben:

Phase 1: Vorbereitung der Analyse

Vor der Analyse des Domänenexperten kann der IT-Experte die für die Analyse verfügbaren Datenquellen im Datenquellen-Repository (K1) bearbeiten. Der IT-Experte erhält hierzu die Möglichkeit weitere Datenquellen zum Repository (K1) hinzuzufügen und Metadaten (beispielsweise Primär-/Fremdschlüssel, Datentypen, ...) über vorhandene Datenquellen anzugeben oder zu ändern. Die hinzugefügten Datenquellen und Metadaten werden anschließend in Phase 2 zur Überwachung der Datenqualität verwendet.

Phase 2: Überwachung während des Analyseprozesses

Die zweite Phase beschreibt die Überwachung der Datenqualität für den Analyseprozesses des Domänenexperten. Die Interaktionen des Domänenexperten mit Komponenten des Prozessmodells (K1-K5) während dieser Phase lassen sich in fünf Abschnitte (S1-S5) unterteilen. Die einzelnen Abschnitte werden im Folgenden beschrieben:

S1 Die zweite Phase beginnt mit der Selektion einer Datenquelle. Der Domänenexperte kann für diese entweder eine bereits im Datenquellen-Repository vorhandene Datenquelle selektieren oder eigenständig eine neue Datenquelle hinzufügen (beispielsweise aus einer Exceldatei). Die selektierte Datenquelle steht ihm im Folgenden zur Verwendung in seinem Analyseprozess zur Verfügung.

S2 Der Datenqualitäts-Service (K4) wird als Nächstes über die Selektion der Datenquelle durch den Domänenexperte informiert. Der Datenqualitäts-Service (K4) verwendet die für die selektierte Datenquelle verfügbaren Metadaten aus dem Datenquellen-Repository (K1) um die im Datenqualitäts-Repository (K2) enthaltenen Maßnahmen zur Auswertung der Datenqualität zu durchsuchen und die Maßnahmen zu identifizieren, welche auf die selektierte Datenquelle anwendbar sind. Die Anwendbarkeit einer Maßnahme aus dem Datenqualitäts-Repository auf die Datenquelle oder auf ein Attribut der Datenquelle kann von mehreren Faktoren abhängen, wie beispielsweise vom Datentypen, von auf den Daten definierten Regeln oder weiteren Metadaten. Ergebnis der Analyse sind anwendbare Maßnahmen zur Überwachung

der Datenqualität sowie Vorschläge für weitere Metadaten, mithilfe welcher die Auswertung der Datenqualität verbessert werden kann. Die Verbesserung kann entweder durch die Optimierung vorhandener Maßnahmen oder die Ermöglichung neuer Maßnahmen erfolgen.

S3 In Schritt 3 kann der Domänenexperte die Metadaten ergänzen, welche zuvor in Schritt 2 als hilfreich für die Auswertung der Datenqualität identifiziert wurden oder die bereits vorhandenen Metadaten anpassen. Für diesen Prozess beinhaltet dass User-Interface (K3) Dialoge zur Ergänzung der Informationen (beispielsweise zur Definition eines Wertebereichs). Der Domänenexperte hat zusätzlich die Möglichkeit die Auswertung der Datenqualität an den aktuellen Kontext seiner Anwendung anzupassen, in etwa durch die Angabe von Prioritäten oder die Verwendung von kontextspezifischen Dimensionen wie die Aktualität (Abschnitt 2.4.1). Dies ermöglicht dem Domänenexperten die Datenqualitätsüberwachung mithilfe seiner Expertise und dem Kontext der Anwendung zu verbessern.

   Der Domänenexperte kann diesen Schritt jederzeit wiederholen und vorhandene Informationen anpassen oder weitere Informationen hinzufügen.

S4 In Schritt 4 werden die Informationen des Domänenexperten verwendet, um ein umfassendes Modell zur Überwachung der Datenqualität zu erstellen. Das Modell setzt sich aus allen anwendbaren Maßnahmen zur Überwachung der Datenqualität mit den zur Anwendung benötigten Informationen zusammen. Das resultierende Modell wird von der Evaluationsumgebung verwendet, um die Datenqualität im Workflow des Domänenexperten zu ermitteln und bei Bedarf automatisierte Korrektur-Vorschläge zu generieren.

S5 Die Evaluationsumgebung liefert ein Feedback über die Datenqualität an verschiedenen Stellen im Analyseprozess des Domänenexperten. Dieses wird direkt in den Analyseprozess integriert, um den Domänenexperten aktiv auf die Qualität der Daten aufmerksam zu machen. Dies ermöglicht dem Domänenexperten eine präzise Einschätzung der Datenqualität an verschiedenen Stellen innerhalb seines Analyseprozesses. Datenqualitätsprobleme können so im Verlauf des Analyseprozesses zurückverfolgt werden.

   Neben einem Feedback zur Datenqualität generiert die Evaluationsumgebung automatisch ausführbare Korrekturvorschläge für gefundene Datenqualitätsprobleme. Der Domänenexperte kann die Korrekturvorschläge nutzen, um Probleme in den Daten zu beheben und die Datenqualität zu verbessern.

Für das Prozessmodell ergeben sich aus diesem Ablauf folgende Eigenschaften im Hinblick auf die Anforderungen des Domänenexperten (Abschnitt 1.1):

Das Prozessmodell (Abbildung 3.1) implementiert die Datenqualitätsüberwachung in den vollständigen Analyseprozess des Domänenexperten (Abschnitt 1.1, A1). Der IT-Experte und der Domänenexperte sind als eigenständige Akteure in den Prozess der Datenqualitätsüberwachung integriert. Für die Auswertung der Datenqualität kann daher auf die kombinierte Expertise beider Akteure zurückgegriffen werden (A3 Abschnitt 1.1). Das Prozessmodell ermöglicht eine automatisierte Ermittlung der Datenqualität (A5 Abschnitt 1.1) und bezieht den Domänenexperten interaktiv in den Prozess der Datenqualitätsüberwachung mit ein (A6 Abschnitt 1.1). Optionen zur Korrektur gefundener Datenqualitätsprobleme werden im Prozessmodell berücksichtigt (A7 Abschnitt 1.1).

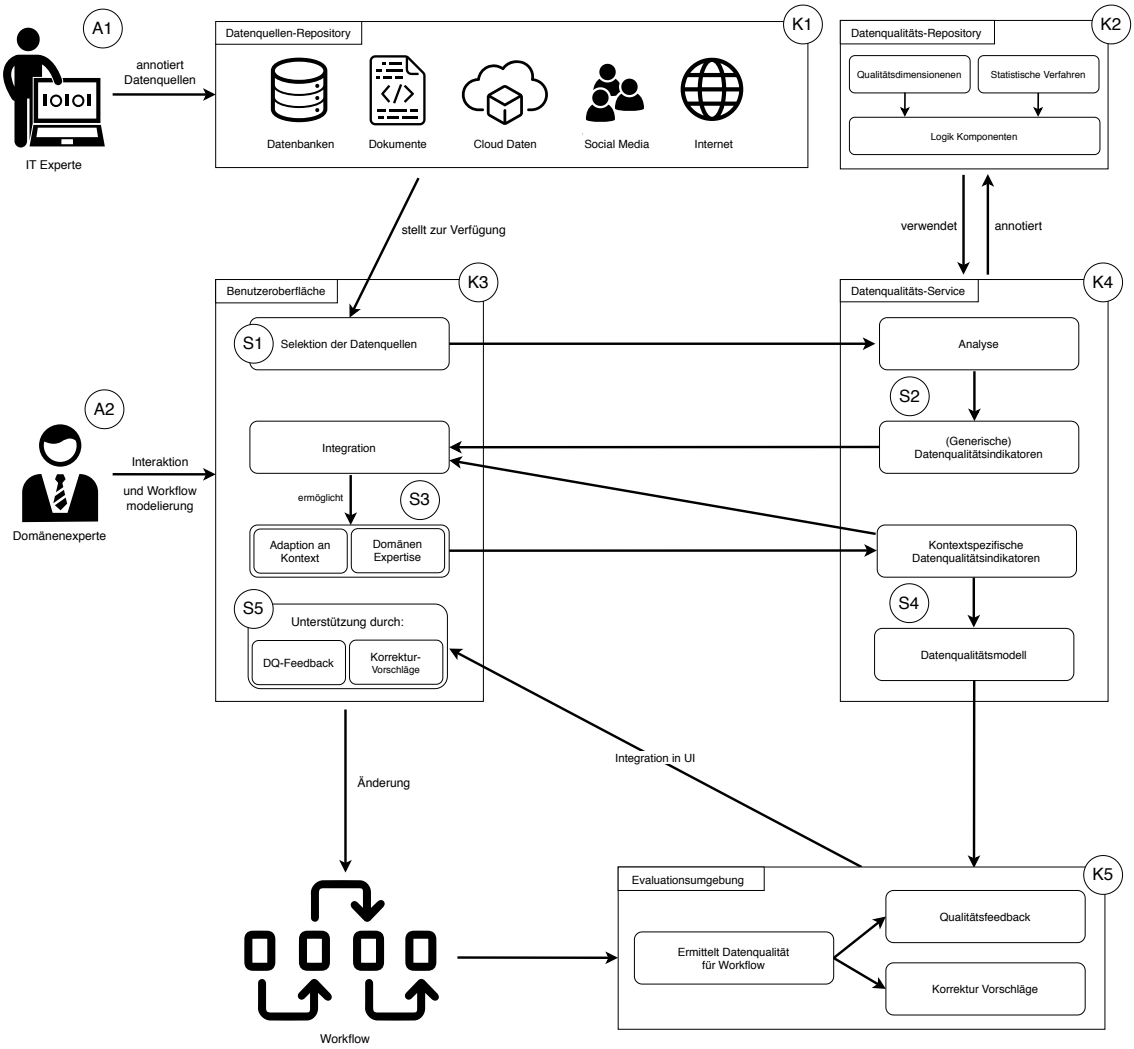

Abbildung 3.1: Prozessmodell zur Integration der Datenqualitätsüberwachung in den Analyseprozess eines Domänenexperten mit den Akteuren A1-A2, den Komponenten K1-K5 und den Schritten S1-S5 (Bildquellen[1])

[1] database by Shmidt Sergey from the Noun Project
Programmer (#637220) from The Noun Project
Account Manager (#551682) by Aneeque Ahmed from The Noun Project
XML by ArmOkay from the Noun Project
cloud data by vecco from the Noun Project
Social by Hrbon from the Noun Project
internet by agus raharjo from the Noun Project
Workflow by Delwar Hossain from the Noun Project

3.1.2 Beispielszenario

Im Folgenden wird die Datenqualitätsüberwachung mithilfe des Prozessmodells anhand eines Beispielszenarios dargestellt. Für das Szenario wird der Maschinenbauingenieur aus Abschnitt 1.1 herangezogen, welcher die Leistungsdaten von Maschinen aus mehreren Fabriken analysieren möchte.

Phase 1: Vorbereitung der Analyse

Der Domänenexperte informiert die IT-Experten der einzelnen Fabriken darüber, dass er die Leistungsdaten der Fabriken für seine Datenanalyse benötigt. Die IT-Experten der verschiedenen Fabriken stellen ihm daraufhin die Leistungsdaten der Maschinen aus ihrer Fabrik (exemplarisch in Tabelle 3.1) im Datenquellen-Repository (K1) zur Verfügung.

Die IT-Experten können hierbei ihre Expertise über die technischen Aspekte der Daten (Schema, Datentypen, ...) nutzen, um die spätere Ermittlung der Datenqualität zu verbessern, indem sie die Daten um Metadaten ergänzen. Die IT-Experten registrieren in diesem Szenario beispielsweise für die Leistungsdaten aus Tabelle 3.1 die Attribute Task- und Maschinen-ID als Fremdschlüssel im System und verlinken den dazugehörigen Datensatz auf welche diese verweisen (hier Daten über die Verfügbaren Maschinen und Daten über die ausgeführten Aufträge). Mithilfe dieser Information kann die Korrektheit der Werte dieser Attribute vom System besser überprüft werden.

Phase 2: Überwachung während des Analyseprozesses

S1 Der Maschinenbauingenieur selektiert die für seine Analyse relevanten Leistungsdaten der Maschinen aus dem Datenquellen-Repository. Im Folgenden werden exemplarisch die Daten aus Tabelle 3.1 verwendet.

S2 Der Datenqualitäts-Service (K4) erhält die vom Maschinenbauingenieur selektierten Datenquellen (hier Tabelle 3.1). Die Datenquelle wurde in Phase 1 bereits von einem IT-Experten mit Informationen über Fremdschlüssel annotiert. Es wird nun unter Berücksichtigung der vorhandenen Metadaten das Datenqualitäts-Repository nach anwendbaren Datenqualitätsmetriken durchsucht. Die Suche nach anwendbaren Maßnahmen zur Überwachung der Datenqualität resultiert daher nicht nur in generischen Maßnahmen (beispielsweise die Analyse der vorhandenen NULL-Werte), sondern auch in Maßnahmen, welche die Informationen aus Phase 1 zur Ermittlung der Datenqualität verwenden. Durch die Information des IT-Experten aus Phase 1, dass es sich im Datensatz Tabelle 3.1 bei den Attributen Task- und Maschinen ID um Fremdschlüssel handelt, können in diesem Schritt bereits Maßnahmen zur Identifikation inkorrekter Werte in den Daten angewandt werden. Ein Beispiel hierfür stellt der Wert Task ID=„R07VV“ aus Tabelle 3.1 dar, welcher als inkorrekt identifiziert werden kann, da die Task ID nicht im Datensatz der ausgeführten Aufträge auf, welche der Fremdschlüssel Task ID verweist vorhanden ist.

   Zusätzlich wird das Datenqualitäts-Repository (K2) nach weiteren Metadaten, die zur Verbesserung der Datenqualitäts-Evaluation genutzt werden können, durchsucht. Im derzeitigen Szenario befindet sich im Datenqualitäts-Repository exemplarisch eine Maßnahme, welche die Konsistenzdimension (Abschnitt 2.4.1) der Daten auswertet. Diese Maßnahme kann zur

Auswertung der Konsistenz beispielsweise die auf einer Datenquelle definierte Regeln der Form „statement 1“ –> „statement“ und die Angabe von Wertebereichen für einzelne Attribute nutzen. Die Angabe dieser Informationen kann daher die Auswertung der Datenqualität verbessern.

Maschine	Typ	Datum	RPM	Laser (W)	Temp (C°)	Status	Task ID
24	FalsFlow 6000	2019-11-18 10:12:00	null	9340	32	working	0XP4X
5	O2-SS	2019-11-18 10:12:00	8796	null	52	working	R07VV
24	FalsFlow 6000	2019-11-18 10:12:30	790	9300	33	working	0XP4X
5	O2-SS	2019-11-18 10:12:30	14000	null	49.9999	working	A1KZ8
24	FalsFlow 6000	2009-11-18 10:13:00	null	0	30	idle	0XP4X

Tabelle 3.1: Exemplarische Überwachungsdaten einer Lasermaschine (FalsFlow) und einer Fräsmaschine (VF2) mit grau hervorgehobenen Fehlern des Beispielszenarios (Abschnitt 3.1.2)

S3 Die in S2 gefundenen Maßnahmen werden dem Maschinenbauingenieur zusammen mit Dialogen zur Ergänzung nützlicher Metadaten angezeigt. Diese Dialoge beinhalten aufgrund der fehlenden Metadaten für die Auswertung der Konsistenzdimension einen Dialog zur Definition dieser. Für die Konsistenzdimension setzen sich die verwendbaren Metadaten hier exemplarisch aus Regeln und Wertebereichen für die Daten zusammen.

Der Maschinenbauingenieur verwendet jetzt seine Expertise über die Daten, um diese Informationen zu ergänzen. Der Maschinenbauingenieur kennt sich beispielsweise mit der Fräsmaschine vom Typ O2-SS aus und weiß, dass diese maximal 12.000 RPM erreichen kann. Er erstellt daher die Regel der Form Typ=„O2-SS“ –> RPM<=„12000“. Informationen können auch später jederzeit ergänzt und verändert werden. Dem Maschinenbauingenieur fällt beispielsweise später ein, dass der Laser FalsFlow 6000 keine drehbaren Teile besitzt und die RPM Angabe daher immer „leer“ sein sollte. Er ergänzt deshalb nachträglich die entsprechende Regel Typ=„FalsFlow 6000“ –> RPM=„null“.

Zusätzlich zu der Ergänzung von Informationen über die Daten kann der Maschinenbauingenieur die Ermittlung der Datenqualität an den Kontext seiner aktuellen Aufgabe anpassen. In diesem Szenario versucht er Rückschlüsse aus den Leistungsdaten der Maschinen zu gewinnen. Die Aktualität der betrachteten Daten ist für ihn daher sehr wichtig, da alte Leistungsdaten keinen Rückschluss auf den aktuellen Zustand der Produktion zulassen. Der Maschinenbauingenieur passt aus diesem Grund die Gewichtung der Aktualitätsdimension entsprechend an,

sodass diese bei der Ermittlung der Datenqualität überproportional berücksichtigt wird. Er gibt außerdem für die Auswertung der Aktualitätsdimension einen entsprechend schnellen Verfall der Daten an.

S4 Aus allen zu diesem Zeitpunkt verfügbaren Informationen über die Daten wird vom Datenqualitäts-Service ein Modell zur Auswertung der Datenqualität generiert. Dieses beinhaltet alle anwendbaren Maßnahmen zur Ermittlung der Datenqualität sowie die von diesen Maßnahmen benötigten Informationen. In diesem Szenario beinhaltet die Auswertung exemplarisch die Überprüfung der Regel Typ=„O2-SS“ –> RPM<=„12000“, die Bestimmung der Aktualität, sowie die Überprüfung der Fremdschlüssel.

Für den Fall, dass sich im späteren Verlauf Informationen über die Daten ändern oder neue hinzukommen, muss dieses Modell überarbeitet werden. Dies tritt beispielsweise bei der nachträglichen Ergänzung der Regel Typ=„FalsFlow 6000“ –> RPM=„null“ durch den Maschinenbauingenieur in S3 auf.

S5 Die Evaluationsumgebung wendet jetzt das Modell auf den Workflow des Maschinenbauingenieurs an, um Datenqualitätsprobleme zu identifizieren. In Tabelle 3.1 sind gefundene Datenqualitätsprobleme (grau) hervorgehoben.

Der Wert „R07VV“ des Attributes „Task ID“ stellt eine Verletzung der Referentiellen Integrität dar (Wert existiert nicht in der referenzierten Datenbank Abschnitt 2.4.1). Für die Erkennung des Problems wurde die Information des IT-Experten, dass es sich bei dem Attribut „Task ID“ um einen Fremdschlüssel handelt, verwendet. Die beiden „RPM“ Werte „790“ und „14000“ verletzen die in Schritt drei (S3) durch den Domänenexperten erstellten Regeln. Die letzte Spalte der Tabelle verletzt das vom Domänenexperten festgelegte Aktualitäts-Kriterium.

Das Ergebnis wird verwendet, um dem Maschinenbauingenieur im User-Interface ein Feedback zur Datenqualität zu geben. Zusätzlich werden Verbesserungsvorschläge für den Maschinenbauingenieur generiert, um die Datenqualität zu verbessern. Ein denkbarer Korrekturvorschlag in diesem Szenario ist die automatische Änderung aller „RPM“ Attribut-Werte auf „null“ für Maschinen des Typs “FalsFlow 6000“.

3.2 Datenquellen-Repository

Das Datenquellen-Repository (Abbildung 3.1, K1) verwaltet die Datenquellen, welche für Analyseprozesse der Domänenexperten genutzt werden können. Die Verwaltung der Datenquellen beinhaltet die Bearbeitung der Datenquellen (hinzufügen, entfernen, anpassen) und die Speicherung zugehöriger Metadaten. Die im Datenquellen-Repository vorhandenen Datenquellen und Metadaten können von IT-Experten und Domänenexperten bearbeitet werden.

3.2.1 Funktionsweise

Die im Datenquellen-Repository vorhandenen Informationen werden durch die kombinierte Expertise des IT-Experten und des Domänenexperten generiert.

Der IT-Experte kann unabhängig vom Domänenexperten Datenquellen zum Datenquellen-Repository hinzufügen und Metadaten zu Datenquellen bereitstellen. Aufgrund seiner Expertise kann der IT-Experte dem Datenquellen-Repositroy auch technisch komplexe Datenquellen (Hadoop, SQL Server, AWS Buckets,...) und Metadaten hinzufügen. Es ist außerdem denkbar, dass der IT-Experte sich besser über die in einer Institution oder in einem Gebiet vorhandenen Datenquellen auskennt und daher in der Lage ist, eine größere Anzahl an relevanten Datenquellen für die Analyse zu identifizieren.

Der Domänenexperte kann während eines Analyseprozesses auf die vom IT-Experten sowie auf die aus früheren Analysen im Datenquellen-Repository vorhandenen Datenquellen und Metadaten zurückgreifen. Dies ermöglicht die Verwendung der vom IT-Experten bereits konfigurierten technisch komplexen Datenquellen (Hadoop, SQL Server, AWS Buckets,...) und Metadaten (Datentypen, Fremdschlüssel,...) im Analyseprozess durch den Domänenexperten. Der Domänenexperte kann bei Bedarf weniger komplexe Datenquellen (beispielsweise lokale Dateien wie Excel, JSON,...) selbständig zum Datenquellen-Repository hinzufügen. Abhängig von der Art der Datenquelle müssen die Daten nicht physikalisch im Datenquellen-Repository enthalten sein, sondern können auch auf externe Ressourcen wie Cloud-Speicher verweisen.

3.2.2 Metadaten

Metadaten, die durch den IT-Experten oder den Domänenexperten (während des Analyseprozesses) zu einer Datenquelle hinzugefügt werden, werden im Datenquellen-Repository gespeichert.

Bei der Speicherung der Metadaten muss zwischen kontextunabhängigen (beispielsweise Datentypen) und kontextspezifischen Eigenschaften der Datenqualität (beispielsweise der Aktualität) unterschieden werden [BS06, Kapitel 2]. Kontextunabhängige Eigenschaften hängen nur von der Datenquelle ab [BS06, Kapitel 2]. Es reicht folglich aus, für jede Datenquelle jeweils alle bekannten kontextunabhängigen Eigenschaften zu speichern. Bei kontextspezifischen Eigenschaften hängt die Eigenschaft auch vom Anwendungsfall des Domänenexperten ab. Kontextspezifische Eigenschaften werden daher in Verbindung mit einer Datenquelle und dem ursprünglichen Analyseworkflow gespeichert.

Während des Analyseprozesses können aus vorhandenen Datenquellen neue Datenquellen mit anderen Eigenschaften entstehen (etwa durch das Zusammenfügen mehrerer Datenquellen). Die Evaluationsumgebung versucht in diesem Fall die Metadaten der alten Datenquellen auf die neue zu übertragen. Bei diesem Prozess kann es zu Konflikten zwischen den zu kombinierenden Metadaten kommen, welche durch den Domänenexperten gelöst werden müssen. Die Metadaten für diese Datenquellen werden daher ebenfalls im Datenquellen-Repository gespeichert, um dem Domänenexperten ein erneutes Lösen dieser Konflikte zukünftig zu ersparen.

3.3 Datenqualitäts-Repository

Das Datenqualitäts-Repository (Abbildung 3.1, K2) enthält Komponenten zur Überwachung der Datenqualität. Diese lassen sich in Komponenten zur Auswertung einzelner Datenqualitäts-Dimensionen und in Komponenten mit deskriptiven statistischen Verfahren unterteilen. Die standardmäßig im Datenqualitäts-Repository enthaltenen Komponenten werten die wichtigsten Dimensionen

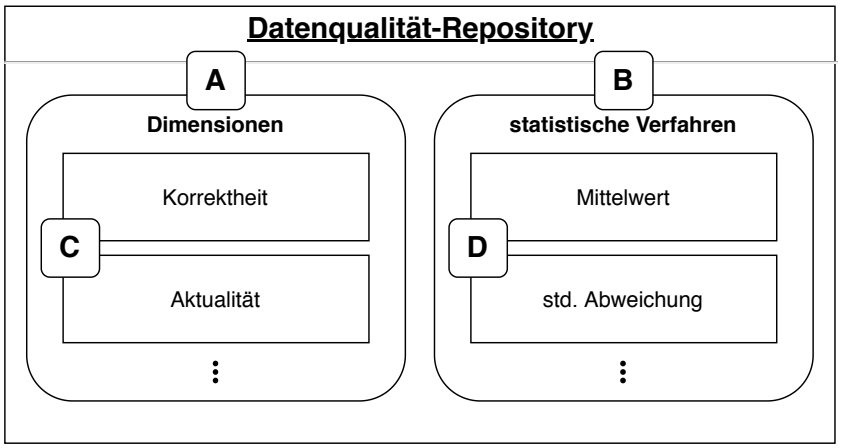

Abbildung 3.2: Aufbau des Datenqualitäts-Repositories mit Unterteilung in Dimensionen (A) und statistische Verfahren (B) und jeweils exemplarischen Komponenten (C, D)

der Datenqualität (aus Abschnitt 2.4.1) und eine Bandbreite statistischer Verfahren aus, um so eine umfassende Beurteilung der Datenqualität zu ermöglichen. Die Auswertung der Datenqualität kann durch die Modifikation der enthaltenen Komponenten jederzeit angepasst werden.

3.3.1 Aufbau des Datenqualitäts-Repository

Das Datenqualitäts-Repository (Abbildung 3.2) ist modular aus eigenständigen Komponenten aufgebaut. Bei den Komponenten wird zwischen statistischen Komponenten und Komponenten zur Auswertung einzelner Dimensionen unterschieden (A, B). Jede ausgewertete Dimension und jedes im Repository vorhandene statistische Verfahren wird daher durch eine eigenständige Komponente (C, D) repräsentiert.

Der Aufbau einer solchen Komponente zur Evaluation einer Datenqualitäts-Dimension setzt sich aus den folgenden vier Bestandteilen zusammen:

1. *Auswertung*
   In jeder Komponente muss die Auswertung der jeweiligen Dimension auf den Datenquellen definiert sein. Die Auswertung ist auf mehreren Granularitätsebenen (Datensatz, Attribute, Tupel, Werte) der Daten definiert. Dadurch wird dem Domänenexperten eine Rückmeldung zu der Datenqualität auf den verschiedenen Granularitätsebenen geboten.

2. *Korrekturoptionen*
   Für die während der Auswertung identifizierten Probleme muss die Komponente (sofern möglich) Korrekturoptionen anbieten. Diese können später vom Domänenexperten für eine automatische Korrektur von Datenqualitätsproblemen (Abschnitt 1.1, A7) verwendet werden. Die Anwendung einer Korrekturmaßnahme bezieht sich dabei stets auf alle Instanzen des Problems in den Daten, da die Korrektur einzelner Problemstellen bei großen Datenmengen nicht zielführend ist.

   Eine Korrekturoption beinhaltet vier wesentliche Bestandteile:

Ausmaß des Problems: Informationen über das Ausmaß eines Problems sind wichtig, um die Bedeutung der Korrektur des Problems einordnen zu können. Der Domänenexperte kann mithilfe des Ausmaßes von Problemen gezielt die wichtigsten Probleme beheben. Zur Quantifizierung kann beispielsweise die Fehlerhäufigkeit oder der Einfluss der Korrektur auf die gesamte Datenqualität verwendet werden.

Maßnahmen zur Problembehebung: Problemspezifische Maßnahmen zur Behebung gefundener Datenqualitätsprobleme. Der Domänenexperte kann für gefundene Datenqualitätsprobleme eine der zur Verfügung stehenden Korrekturmaßnahmen für eine automatisierte Korrektur des Problems auswählen.

Prädiktive Auswirkungsanalyse: Eine Vorhersage zu den Auswirkungen der Anwendung der Korrekturmaßnahmen auf die Datenqualität.

UI-Elemente: Um die spätere Darstellung der Korrekturoption in der Benutzeroberfläche zu ermöglichen. Diese umfassen beispielsweise textuelle und grafische Beschreibungen von Problemen und Korrekturmaßnahmen.

3. *Identifikation zur Ausführung benötigter Parameter*
   Um die Expertise des Domänenexperten effizient nutzen zu können müssen die Metadaten identifiziert werden, welche für die Überwachung der Datenqualität notwendig sind oder diese unterstützen. Nur so kann der Domänenexperte gezielt nach diesen Informationen gefragt werden. Jede Komponente muss daher klare Anforderungen an die zur Ausführung benötigten Parameter beinhalten. Durch einen einfachen Abgleich der vorhandenen Metadaten mit den zur Ausführung der Komponente benötigten Parametern können fehlende Informationen identifiziert werden. Diese können dann durch den Domänenexperten zur Verbesserung der Datenqualitätsüberwachung ergänzt werden.

   Parameter können in zur Auswertung der Dimension zwingend benötigte Parameter und optionale (die Auswertung verbessernde) Parameter unterteilt werden.

4. *UI-Elemente*
   Für die Ergänzung von Informationen, die Anzeige vorhandener Informationen und die Darstellung von Auswertungsergebnissen zu ermöglichen werden UI-Elemente benötigt. Um die modulare Änderung einzelner Komponenten gewährleisten zu können, müssen UI-Elemente zusammen mit den zugehörigen Komponenten gespeichert werden. Andernfalls muss bei der Änderung einer Komponente nicht nur die Komponente selbst sondern auch die Benutzeroberfläche angepasst werden.

Das Ziel dieser Komponenten ist die Beschreibung der Daten [Nic07], damit der Domänenexperte ein besseres Verständnis für diese entwickeln kann. Für den Aufbau der Komponenten deskriptiver statistischer Verfahren werden weniger Bestandteile als für Komponenten, welche Datenqualitäts-Dimensionen auswerten, benötigt. Weil die statistischen Komponenten die Eigenschaften der Daten beschreiben und keine konkreten Problemstellen der Daten identifizieren, benötigen sie auch keinen Bestandteil zur Korrektur gefundener Problemstellen. Der Bestandteil „Korrekturoptionen“ kann daher für statistische Komponenten vernachlässigt werden. Des Weiteren kann der Bestandteil „Identifikation zur Ausführung benötigter Parameter“ für statistische Verfahren stark vereinfacht werden, da für die Anwendung deskriptiver statistischer Maße lediglich zwischen kategorischen (Klassifikationen) und nummerischen (Kardinal-skaliert) Daten unterschieden wird [Nic07]. Es

werden keine weiteren Metadaten zur Anwendung von statistischen Komponenten benötigt [Nic07]. Die Bestandteile zur Auswertung der statistischen Maße und die UI-Elemente gestalten sich analog zu Komponenten, die Datenqualitäts-Dimension auswerten.

Durch diesen modularen Aufbau kann die Datenqualitätsüberwachung angepasst werden. Die Anpassung der Datenqualitätsüberwachung erfolgt über das Hinzufügen neuer Komponenten oder die Änderung bereits vorhandener Komponenten. Dies ermöglicht die Anpassung der Datenqualitätsüberwachung an die Bedürfnisse spezifischer Anwendungsfälle. Es kann beispielsweise ein für einen bestimmten Anwendungsfall trainiertes Maschine-Learning Modell bei der Auswertung einer Dimension verwendet werden. Hierzu muss nur die für die Dimension zuständige Komponente angepasst werden. Da Änderungen im Datenqualitäts-Repository Programmierkenntnisse erfordern müssen diese vom IT-Experten ausgeführt werden. Die Anpassung des modularen Datenqualitäts-Repository bietet dem IT-Experten eine weitere Möglichkeit, sich am Analyseprozess zu beteiligen (Abschnitt 1.1, A3).

3.3.2 Metriken für Komponenten

Für die Auswertung der Datenqualität benötigen die Komponenten Metriken zur Quantifizierung der gewünschten Dimension der Datenqualität (Abschnitt 2.4.2). Die Auswahl einer geeigneten Metrik kann anhand der Konformität der Metrik zu den Anforderungen an Qualitätsmetriken (Abschnitt 2.4.2) getroffen werden.

Bei der Auswahl von Metriken anhand der Anforderungen (Abschnitt 2.4.2) ergeben sich aus den Anforderungen des Domänenexperten (Abschnitt 1.1) und dem Prozessmodell (Abschnitt 3.1) verschiedene Prioritäten für die verschiedenen Anforderungen an die Qualitätsmetriken. Die Priorität der einzelnen Anforderungen wird im Folgenden kurz erläutert:

Prioritäten der Anforderungen an Qualitätsmetriken

Die für die Unterstützung des Domänenexperten wichtigsten Anforderungen an Metriken sind die Anforderungen A1 (Minimum und Maximum) und A2 (Intervallskaliert). Aus der Erfüllung dieser Anforderungen resultieren die folgenden Eigenschaften, die für die Unterstützung des Domänenexperten benötigt werden [HHK+18]:

Vergleichbarkeit von Änderungen
Die Vergleichbarkeit von Änderungen der Metrikwerte ist nur bei Erfüllung der beiden Anforderungen (A1, A2) gegeben. Der Domänenexperte benötigt diese Eigenschaft während des Analyseprozesses um die Auswirkung von Operationen und angewendeten Korrekturmaßnahmen auf die Datenqualität einschätzen zu können [HHK+18].

Interpretierbarkeit der Werte
Die Existenz eines Minimums und eines Maximums (A1) und die Intervallskalierung (A2) ermöglichen die einfache Interpretation der Metrikwerte. Da der Domänenexperte keine Expertise auf dem Bereich der Datenqualität besitzt, ist es wichtig, dass er die Ergebniswerte möglichst einfach verstehen kann [HHK+18].

Vereinheitlichung der Werte
Die Metriken, welche die Anforderungen A1 und A2 erfüllen, können durch Transformation der Werte rationalskaliert (Bruchteil des maximalen Wertes) abgebildet werden. Dies ermöglicht eine normierte Darstellung der Metrikwerte (als Intervall [0-1]). Dadurch können die Auswertungsergebnisse der verschiedenen Dimensionen einheitlich dargestellt werden [HHK+18].

Des Weiteren ist für die Unterstützung des Domänenexperten auf verschiedenen Granularitätsebenen (Abschnitt 1.1, A2) die Aggregation der Metrikwerte auf die verschiedenen Granularitätsebenen der Daten relevant. Die Anforderung „Aggregation der Werte“ (Abschnitt 2.4.2, A4) an die Werte der Metrik ist daher von hoher Bedeutung.

Während die Erfüllung der Anforderungen an Qualitätsmetriken A3 und A5 auch wünschenswert ist, ist eine Verletzung dieser weniger kritisch im Hinblick auf die Unterstützung der Domänenexperten während des Analyseprozesses.

Ordnungsrahmen

Damit sichergestellt wird, dass die Metriken, welche zur Auswertung der Datenqualität verwendet werden die für die Unterstützung des Domänenexperten wichtigen Anforderungen erfüllen, wurden die Metriken aus Abschnitt 2.5 den Anforderungen an Datenqualitätsmetriken gegenübergestellt:

Aktualität

Metrik:	Probabilistisch	Zeitlich begrenzt	Update-Häufigkeit
A1	✗	✓	✗
A2	✗	✗	✗
A3	✓	(✗)*	(✓)**
A4	✓	✓	✓
A5	Kontextabhängig***	Kontextabhängig***	Kontextabhängig***

* nur für Werte mit festem „shelf life“
** wenn Upd(A) in der Realen Wert existiert
*** wenn das Alter der Werte in der Realen Welt effizient bestimmt werden kann

Tabelle 3.2: Auswertung der Metriken für die Aktualitätsdimension (Gegenüberstellung des Zeitlich begrenzten Ansatzes aus Heinrich et al. [HHK+18])

Der Gemischte Ansatz zur Auswertung der Aktualität aus Abschnitt 2.5 nimmt entweder die Charakteristiken des Probabilistischen oder die des Zeitlich begrenzten Ansatzes an.
Für die Messung der Aktualitätsdimension für Attribute mit über die Zeit abnehmender Qualität ist die Anforderungen A1 ohne die Annahme einer Schwelle ab welcher die Datenqualität unverändert bleibt unerfüllbar. Für die Erfüllung der Anforderung A1 benötigt eine Metrik ein Minimum. Dieses kann ohne Annahme einer solchen Schwelle nicht existieren, da sich die Qualität der Werte ansonsten in jedem Zeitschrift weiter verschlechtert.

Vollständigkeit

Metrik:	Verhältnis fehlender Werte	Verhältnis Tupel
A1	✓	✓
A2	✓	✓
A3	✓	✓
A4	✓	✗*
A5	✓	✓

* Für die Unterstützung des Domänenexperten ist die Aggregation der Metrikwerte auf die Ebene einzelner Attribute relevant. Diese Aggregation wurde von Heinrich et al. nicht betrachtet [HHK+18].

Tabelle 3.3: Auswertung der Metriken für die Vollständigkeitsdimension (Gegenüberstellung des auf dem Verhältnis der Tupel basierendem Ansatzes aus Heinrich et al. [HHK+18])

Korrektheit

Metrik:	Verhältniss Korr. zu Inkorr. Werte (2.5)	Mit Abstandsfuntkion (2.5)
A1	✓	✗
A2	✓	✗
A3	✓	✓
A4	✓	✓*
A5	(✗)**	(✗)**

* Aggregation aus [Hin02] kann verwendet werden. Zur Erfüllung muss jedoch für die Aggregation auf Datenbankebene eine gewichtete Summe verwendet werden [HHK+18].
** Abhängig von der Verfügbarkeit von Referenzdaten

Tabelle 3.4: Auswertung der Metriken für die Korrektheitsdimension (Gegenüberstellung des Ansatzes mit Abstandsfunktion aus Heinrich et al. [HHK+18])

A5 ist für die Korrektheitsdimension nur schwer erfüllbar, da Referenzwerte zum Vergleich der Werte oder zur Ermittlung der inkorrekten Werte benötigt werden. Die Ermittlung dieser kann unter Umständen sehr teuer/schwer sein [HHK+18].

Konsistenz

Metrik:	Verhältnis	gewichtete Summe)	max. konfliktfreie Menge)
A1	✓	✗	✓
A2	✓	(✓)*	(✗)**
A3	✓	✓	✓
A4	✓	✗	✗
A5	(✓)	(✓)	✓

* nur für lineare Parameter
** $C_{min}(D)$ nicht linear für variable CFDs (**??**)

Tabelle 3.5: Auswertung der Metriken für die Konsistenzdimension (Gegenüberstellung des Ansatzes der gewichteten Summe aus Heinrich et al. [HHK+18])

3.3.3 Spezifikation grundlegender Komponenten

Die standardmäßig im Datenqualitäts-Repository enthaltenen Komponenten werten die wichtigsten Dimensionen der Datenqualität aus Abschnitt 2.4.1 aus. In diesem Abschnitt wird die Zusammensetzung der einzelnen Komponenten anhand des in Abschnitt 3.3.1 definierten Aufbaus von Komponenten des Datenqualitäts-Repositories beschrieben. Die Auswahl der Metrik zur Evaluation der einzelnen Komponenten richtet sich nach den Prioritäten für die Anforderungen der Datenqualitätsmetriken (Abschnitt 3.3.2).

Korrektheit

Die Korrektheitsdimension beschreibt die Abweichung der Daten zu den Entitäten, welche von diesen modelliert werden (Abschnitt 2.4.1). Die Evaluation der Korrektheitsdimension ist schwierig [MTV19; Los11, Kapitel 8], da für diese zuerst inkorrekte Werte identifiziert werden müssen [ASWW18][Los11, Kapitel 8]. Die Fachliteratur identifiziert hierfür eine Reihe an Methodiken:

- *Abgleich mit korrekten Datenquellen*

  Die Korrektheit von Daten kann durch einen Vergleich der Daten mit korrekten Referenzdaten bestimmt werden. In vielen Fällen sind Referenzdaten jedoch nicht vorhanden oder die Akquisition dieser ist komplex und mit erheblichem Aufwand verbunden [ZZ16; ASWW18; CZ15].

  Mithilfe von „Entity Resolution“[2] können unter Umständen auch semantisch inkorrekte Werte identifiziert werden [ASWW18]. Dies ist der Fall, wenn ein Tupel aus den Daten sich auf dieselbe Entität wie ein Tupel aus den Referenzdaten bezieht [ASWW18]. In diesem Fall stellen die Referenzdaten die für die Entität (also die semantisch) korrekten Werte [ASWW18].

  Für die Unterstützung des Domänenexperten können im Datenquellen-Repository Referenzdaten für die vorhandenen Datenquellen angegeben werden. Die Anschaffung/Auffindung und Verlinkung korrekter Referenzdaten sollte durch den IT-Experten vorgenommen werden, da dieser sich am besten mit den verfügbaren Daten und deren Zustand auskennt.

- *Verwendung von Geschäftsregeln*

  Geschäftsregeln (Regeln der Form col1=„x“, col2=„y“ –> col3=„z“) können zur Identifikation inkorrekter Werte verwendet werden [Gar97; AW14; ASWW18]. Hierzu wird die Konformität der Daten zu vorher definierten Regeln überprüft [Gar97; AW14; ASWW18].

  Die Geschäftregeln gehen aus der Domäne, aus welcher die Daten stammen, hervor. Daher eignet sich der Domänenexperte mit seiner Expertise am besten für die Bereitstellung.

  Eine Möglichkeit Geschäftsregeln zu spezifizieren bietet die Modellierungssprache Decision Model and Notation[3] (DMN), welche mit „Decision Tables“ einen Standard zur Darstellung von Geschäftsregeln spezifiziert [OMG19; VPV+19]. Die Verwendung von DMN bietet eine für den Domänenexperten einfach verständliche und einfach erlernbare Modellierung von

[2] https://doi.org/10.1016/C2009-0-63396-1
[3] https://www.omg.org/dmn/

Geschäftsregeln [OMG19; HDV17]. Die DM Notation ermöglicht außerdem die einfache Verarbeitung der vom Domänenexperten modellierten „Decision Tables“ in Form von XML Objekten [OMG19].

Die DM Notation wird etwa von Valencia et al. [VPV+19] im Hinblick auf Qualitätsdimensionen eingesetzt.

- *Data Mining*

  Mithilfe von Data Mining können automatisiert Regeln (Assoziationsregeln) aus den Daten abgeleitet werden [HMHN07; AW14; Jud15; YP10]. Die resultierenden Regeln können dann analog zu den Geschäftsregeln zur Überprüfung der Daten verwendet werden [HMHN07; AW14; Jud15; YP10]. Data Mining kann zusätzlich zu den vom Domänenexperten spezifizierten Regeln verwendet werden, da bei manueller Definition der Regeln oftmals nicht alle Abhängigkeiten der Daten berücksichtigt werden (etwa wenn der Domänenexperte vergisst, eine Regel zu definierten oder eine Abhängigkeit selbst nicht kennt) [AW14; HMHN07].

  Der Domänenexperte kann analog zu He et al. [HVS+16] (Falcon aus Abschnitt 2.8.1) seine Expertise verwenden um die automatisiert abgeleiteten Regeln zu validieren. Für die Validierung der Regeln können diese ebenfalls in der DM Notation dargestellt werden. Da das System unter Umständen eine hohe Anzahl an Regeln in den Daten findet, ist bei der Validierung eine Priorisierung der Regeln abhängig von ihrer „confidence“ (für eine Regel X–>Y ist die „confidence“ die Wahrscheinlichkeit für das Auftreten von X und Y im Datensatz geteilt durch die Wahrscheinlichkeit von Y [AW14]) sinnvoll. Anstatt der „confidence“ können auch andere Maße für die Priorisierung von Regeln bei der Validierung verwendet werden [TKS04].

- *Machine Learning zur Identifikation von Referenzdaten*

  Machine Learning kann zur automatisierten Identifikation von Referenzdaten in einer Sammlung von Daten verwendet werden [MTV19]. Dadurch können, zu den im Datenquellen-Repository vorhandenen Datenquellen, automatisiert Referenzdaten identifiziert werden. Dies ermöglicht die Verwendung von Referenzdaten auch für Datenquellen, welche vom Domänenexperten erst während der Analyse hinzugefügt werden.

Für die Identifikation von inkorrekten Werten in den vom Domänenexperten verwendeten Daten ist eine Kombination dieser Methodiken möglich. Es müssen nicht zwangsläufig alle Methodiken implementiert werden.

Anschließend an die Identifikation inkorrekter Werte wird für die Evaluation der Korrektheit die Metrik auf Basis des Verhältnisses der korrekten zu den inkorrekten Werten (Abschnitt 2.6) verwendet. Die Auswertung mit dieser Metrik stellt die für die Unterstützung des Domänenexperten wichtigsten Anforderungen (A1, A2, A4) sicher (Abschnitt 3.3.2).

Von der Komponente verwendbare Parameter:

Die von der Komponente zur Auswertung der Korrektheit benötigten Parameter sind abhängig von den implementierten Methoden zur Identifizierung inkorrekter Daten. Nachfolgend wird eine Übersicht über die Parameter, die von dieser Komponente verwendet werden können, mit der jeweiligen Rolle (IT-Experte oder Domänenexperte), welche zur Bereitstellung am besten geeignet ist, gegeben.

- *[Abgleich mit korrekten Datenquellen] Referenzdaten (IT-Experte)* Da der Domänenexperte keine Erfahrung im Umgang mit fehlerhaften Daten hat, besteht das Risiko, dass die von ihm verwendeten Referenzdaten fehlerhaft sind. Der IT-Experte kennt sich außerdem besser mit den verfügbaren Datenquellen aus und eignet sich daher für die Angabe von Referenzdaten am besten.
- *[Machine Learning] (automatisierte Suche)] Sammlung potentieller Referenzdaten-Quellen (IT-Experte)* Der IT-Experte kennt sich am besten mit den verfügbaren Daten aus. Er eignet sich daher am besten für die Bereitstellung einer Sammlung potenzieller Referenzdaten.
- *[Geschäftsregeln] Regeln zur Überwachung (Domänenexperte)* Die Geschäftsregeln spiegeln Zusammenhänge innerhalb der Daten wieder. Diese gehen aus der Domäne, aus welcher die Daten stammen, hervor. Der Domänenexperte eignet sich daher aufgrund seiner Expertise über diese Domäne am besten zur Angabe der Geschäftsregeln.

Die Korrektheit kann bereits ausgewertet werden, sobald Parameter zur Anwendung einer der oben aufgezählten Methoden verfügbar sind. Durch die Bereitstellung weiterer Parameter lässt sich das Auswertungsergebnis jedoch präzisieren. Das Data Mining Verfahren benötigt keine direkten Parameter für die Ableitung der Regeln. Es wird jedoch der Domänenexperte für die Verifikation der Regeln benötigt.

Korrekturoptionen für identifizierte Probleme:

Die für die Korrektheitsdimension erstellbaren Korrekturoptionen sind abhängig von der Methodik, durch welche das Problem identifiziert wurde. Für die verschiedenen Methoden zur Identifizierung inkorrekter Werte werden daher im Folgenden eigenständige Optionen zur Korrektur von Problemen vorgestellt:

- *Regelverletzungen*

  Bei der Verletzung einer Regel (entweder einer Geschäftsregel oder einer durch Data Mining erstellte Assoziationsregel) können mehrere Maßnahmen zur Korrektur eines Problems verwendet werden:

  Werte an Regel angleichen

  Für Regeln, die nur einen korrekten Wert für die Problemstelle zulassen (etwa bei einer Regel der Form Maschinentyp=„Laser“ –> Maschine={„FalsFlow 5000“), können die Werte der Problemstellen auf den korrekten Wert (im Beispiel „FalsFlow 5000“) abgeändert werden. Dies stellt die Konformität der Werte mit der Regel sicher.

  Bei mehreren möglichen Werten (etwa bei einer vom Domänenexperten definierten Regel der Form Maschinentyp=„Laser“ –> Maschine={„FalsFlow 5000“, „FalsLaser Cell 5000“}) ist eine probabilistische Korrektur auf Basis der möglichen korrekten Werte (für das obige Beispiel „FalsFlow 5000“, „FalsLaser Cell 5000“) möglich.

  Regel anpassen

  Im Falle eines korrekten Wertes, welcher von der Regel nicht berücksichtigt wird, muss nicht der Wert selbst, sondern die Regel entsprechend korrigiert werden (entweder durch Hinzufügen des Wertes zur Menge der korrekten Werte oder durch Abänderung eines vorhandenen korrekten Wertes).

Regel löschen

Im Falle einer inkorrekten Regel reicht eine Anpassung der Regel nicht aus. Die Löschung der Regel ist in diesem Fall vorzuziehen.

Betroffene Einträge/Werte löschen

Falls die Anwendung einer probabilistischen Korrektur und die Abänderung der Regel vom Domänenexperten nicht gewollt ist, können die gefundenen Problemstellen gelöscht werden. Für Attribute, welche NULL-Werte enthalten können, kann die Löschung der betroffenen Attributwerte ausreichend sein. Anderenfalls sollte das komplette Tupel, welches das Problem beinhaltet, gelöscht werden, da die Korrektur sonst die Qualität der Daten in der Vollständigkeitsdimension verschlechtert.

- *Vergleich mit Referenzdaten*

  Bei der Verwendung von Referenzdaten zur Identifikation inkorrekter Werte sind die vorhandenen Korrekturmaßnahmen eingeschränkter. Die direkte Korrektur falscher Werte ist in diesem Fall nur dann möglich, wenn Problemstellen in Entitäten, welche in den Referenzdaten vorhanden sind, auftreten. In diesem Fall kann mithilfe von Entity Resolution der korrekte Wert identifiziert werden [ASWW18]. Da die Referenzdaten eine größere Anzahl an potenziell richtigen Werten als einzelne Regeln (alle für das jeweilige Attribut zulässige) enthalten, ist eine probabilistische Korrektur komplexer.

  Die weiteren Maßnahmen zur Korrektur von Problemen beschränken sich daher auf:

  Wert erlauben

  Falls der Wert korrekt ist, aber in den Referenzdaten fehlt, kann der Wert bei der zukünftigen Auswertung als korrekt betrachtet werden.

  Betroffene Einträge/Werte löschen

  Erfolgt analog zur Löschung der Werte bei der Identifikation durch Regelverletzung.

Methodenunabhängig kann außerdem (abhängig von den Daten) eine Korrektur inkorrekter Werte auf Basis der vorhandenen Daten vorgenommen werden. Etwa durch Approximation der Werte mithilfe von Machine Learning [JPR19]. Der Erfolg dieser Korrekturmethode hängt von den Daten (genauer von der Präzision, mit welcher diese vorhergesagt werden können) ab.

Vollständigkeit

Die Vollständigkeit beschreibt, inwiefern das Ausmaß der vorhandenen Daten ausreicht, um eine bestimmte Aufgabe zu erfüllen (Abschnitt 2.4.1).

Die Auswertung der Vollständigkeitsdimension kann unter Annahme der „Closed World Assumption“ (Betrachtung der Vollständigkeit innerhalb der Daten Abschnitt 2.4.1) ohne Berücksichtigung der Semantik der Werte (Grund des Fehlens und Werte mit Bedeutung „NULL“ Abschnitt 2.4.1) anhand der in den Daten verfügbaren Werten mit dem Wert „NULL“ ausgewertet werden [Kai10]. Eine solche Auswertung benötigt keine weiteren Informationen über die Daten. Sind für die Auswertung der Daten keine weiteren Informationen bekannt, stellt sie daher die Basis zur Auswertung der Vollständigkeitsdimension.

Mithilfe von Informationen des Domänenexperten und IT-Experten kann jedoch zur Auswertung auch die „Open World Assumption “ (Miteinbeziehung von Werten außerhalb der Daten Abschnitt 2.4.1) angenommen werden. Batini und Scannapieco[BS06, Kapitel 2.2] schlagen hierfür die Verwendung der Kardinalität der Entitäten der Realwelt zur Bestimmung der nicht im Datensatz enthaltenen Werte vor. Diese kann dann mit den im Datensatz vorhandenen Entitäten verglichen werden, um die Anzahl der fehlenden Entitäten zu identifizieren.

Ein Beispiel kann anhand eines Datensatzes gegeben werden, der Informationen zu den Maschinen einer Fabrik beinhaltet. Der Domänenexperte weiß aus Erfahrung, dass die Fabrik insgesamt 42 Maschinen (Entitäten) besitzt. Beinhaltet der Datensatz nur 34 verschiedene Maschinen kann daraus geschlussfolgert werden, dass er unvollständig (8 Maschinen fehlen) ist.

Die Informationen des Domänenexperten und IT-Experten können außerdem verwendet werden, um die Semantik (Abschnitt 2.4.1) bei der Auswertung zu berücksichtigen. Hierfür können Werte mit der semantischen Bedeutung „NULL“ (beispielsweise „-“, „foo„, „xxx@xxx.de“,…) angegeben werden. Des Weiteren können in der Real-Welt nicht existierende Werte durch Regeln angegeben werden. Ein Beispiel für einen solchen Wert stellt die Lasermaschine „FalsFlow 6000“ aus Tabelle 3.1, welche als Lasermaschine kein Attribut RPM besitzt, dar. Fehlende RPM Werte dieser Maschine sollten nicht zu einer Verschlechterung der Bewertung der Vollständigkeitsdimension führen [Kai10]. Es kann die Regel Typ=„FalsFlow 6000“ –> RPM=NULL angegeben werden, um dies bei der Auswertung entsprechend zu berücksichtigen zu können. Für die Angabe dieser Regeln kann analog zur Angabe der Regeln der Korrektheitskomponente die DM Notation genutzt werden.

Als Metrik zur Auswertung der Vollständigkeit wird das Verhältnis der gesamten Werte zu den fehlenden Werten verwendet (Abschnitt 2.5). Die Auswertung mit dieser Metrik stellt die für die Unterstützung des Domänenexperten wichtigsten Anforderungen (A1, A2, A4) sicher (Abschnitt 3.3.2).

Von der Komponente verwendbare Parameter:

Für die Auswertung der Vollständigkeit lassen sich im obigen Auswertungsprozess mehrere Parameter identifizieren, welche vom IT-Experten und dem Domänenexperten gestellt werden können. Die Vollständigkeit kann auch ohne diese Parameter ausgewertet werden. Die Angabe der Parameter ist optional und dient der Verbesserung des Ergebnisses durch Berücksichtigung mehrerer Faktoren bei der Auswertung.

- *Kardinalitäten (Domänenexperte)*

  Kardinalitäten sind an die Domäne, aus welcher die Daten erhoben wurden, gebunden. Die Angabe von Kardinalitäten sollte daher der Domänenexperte übernehmen.

- *Semantische „NULL“-Werte (Domänenexperte, IT-Experte)*

  In den Daten enthaltene semantische „NULL“-Werte können von beiden Parteien identifiziert und angegeben werden, da es sich bei diesen oftmals um für Menschen offensichtliche Filler-Werte (beispielsweise „-“, „foo“, „xxx@xxx.de“,…) handelt und daher keine besonderen Kenntnisse zur Angabe dieser benötigt werden.

- *Regeln für nicht existierende Werte (Domänenexperte)*

  Damit Werte, die in der Real-Welt nicht existieren, identifiziert werden können, wird eine Expertise über die Domäne, aus welcher die Daten erhoben wurden, benötigt. Für die Angabe von Regeln für nicht existierende Werte eignet sich daher der Domänenexperte am besten.

- *„NULL"-Werte erlauben (Domänenexperte)*

  Für Fälle, in welchen die Beschreibung der in der Real-Welt nicht existierender Werte durch Regeln nicht (ohne erheblichen Aufwand) möglich ist, sollte der Domänenexperte die Option haben, die Überwachung der Vollständigkeit für einzelne Attribute auszusetzen („NULL"-Werte für einzelne Attribute zu erlauben). Ein Beispiel für einen solchen Fall stellt ein Szenario in welchem die Daten aus Tabelle 3.1 exemplarisch mehrere tausende verschiedene Maschinentypen enthalten dar. In einem solchen Fall kann von dem Domänenexperten nicht die Definition der Regeln aller Lasermaschinen (welche keinen RPM besitzen) erwartet werden.

  ist es unrealistisch, den Domänenexperten die Regeln für alle Lasermaschinen die keinen RPM Wert haben definieren zu lassen.

Die Vollständigkeit kann auch ohne diese Parameter ausgewertet werden. Die Angabe der Parameter ist optional und dient ausschließlich der Verbesserung des Ergebnisses durch die Berücksichtigung mehrerer Faktoren bei der Auswertung.

Korrekturoptionen für identifizierte Probleme:

Da dem System nur die Informationen „Wert fehlt" vorliegen, ist eine Korrektur mehrerer Werte auf einmal schwierig, da sie nicht durch eine gemeinsame Eigenschaft (beispielsweise bei der Korrektheit den gleichen inkorrekten Wert) miteinander gruppiert werden können. Die separate Korrektur einzelner Fehler ist für große Datenmengen jedoch nicht zielführend. Es wird daher bei den nachfolgenden Korrekturmaßnahmen versucht die Abhängigkeit von Problemen zu anderen Attributen des Datensatzes zu nutzen, um mehrere Problemstellen auf einmal zu beheben.

- *Regel(n) erstellen*

  Existiert ein Wert in der Real-Welt nicht, so kann das Problem durch die automatisierte Erstellung der entsprechenden Regel(n) behoben werden. Dem Domänenexperten werden für die automatisierte Erstellung von Regel(n) Werte anderer Attribute des Datensatzes gezeigt, welche häufig mit dem Wert NULL des Attributes, für welches die Regel(n) erstellt werden sollen, auftritt. Vom Domänenexperten werden aus diesen Werten alle Werte ausgewählt, aus welchen die Nichtexistenz des Wertes in der Real-Wert hervorgeht. Aus den ausgewählten Werten können anschließend automatisiert Regel(n) erstellt werden, um die nicht Existenz der NULL-Werte in der Real-Welt zu berücksichtigen. Die Regeln haben die Form Attribut1=„ausgewählter Wert" –> ProblemAttribut=NULL.

  Durch die Erstellung der Regeln können mehrere Probleme auf einmal behoben werden.

- *Tupel entfernen*

  Tupel mit fehlendem Wert abhängig von Attributwerten löschen. Falls eine Kardinalität angegeben ist, verschlechtert diese Maßnahme (unter Umständen) die Vollständigkeit des Datensatzes (da auch vorhandene Werte anderer Attribute gelöscht werden) und sollte dem Domänenexperten daher nicht vorgeschlagen werden.

Zusätzlich ist die Korrektur der fehlenden Werte durch Approximation auf Basis der vorhandenen Daten durch die Verwendung von Maschine Learning möglich [JPR19].

Konsistenz

Die Konsistenz beschreibt das Ausmaß, zu welchem Widersprüche innerhalb der Daten enthalten sind (Abschnitt 2.4.1).

Für die Anwendung einer Metrik zur Quantifizierung der Konsistenzdimension müssen zuerst Widersprüche in den Daten identifiziert werden (Abschnitt 2.5). Es kann zwischen verschiedenen Arten von Widersprüchen unterschieden werden (Abschnitt 2.5):

- *Unterschiedliche Formate*

  Widersprüchliche Formate eines Attributes (beispielsweise „01.04/2020“ und 06.07.20202“) verletzen die Repräsentative Konsistenz [BM09; LPFW06, Kapitel 4]. Die Überprüfung von Formaten muss allerdings nicht bei jedem Attribut angewandt werden [LPFW06, Kapitel 4]. Für das Konzept wird daher der IT-Experte und der Domänenexperte verwendet, um Formatvorlagen für Attribute, bei denen das Format beachtet werden soll, anzugeben.

- *Semantische Widersprüche*

  Semantische Widersprüche gehen aus der Semantik der Daten hervor [HBSA18; BM09; LPFW06]. Sie überschneiden sich mit dem semantischen Aspekt der Korrektheit. Sie können daher ebenfalls mithilfe von Regeln identifiziert werden [HBSA18].

  Es kann darüber nachgedacht werden, den semantischen Aspekt der Konsistenzdimension oder den semantischen Aspekt der Korrektheitsdimension zu vernachlässigen, um Überschneidungen zwischen den Dimensionen zu vermeiden.

- *Verletzung von Entitäts Regeln*

  Die Entitäts Regeln setzen die Eindeutigkeit aller Primärschlüssel voraus und fordern einen „NOT NULL“ Wert für alle Attribute die Bestandteile eines Primärschlüssels sind [LPW04; **.**]

  Da ein Anwender ohne technische Expertise die Rolle eines Primärschlüssels nicht einordnen kann, muss dieser vom IT-Experten gestellt werden. Für den Domänenexperten können die Eigenschaften von Primärschlüsseln (keine Duplikate, NOT NULL-Wert) als Option zur Angabe von Metadaten integriert werden.

- *Verletzung von Referentiellen Integritätsregeln*

  Eine Verletzung der Referentiellen Integritätsregeln liegt vor, wenn einem Fremdschlüssel kein Primärschlüssel in den Daten, auf welche er verweist, zugeordnet werden kann [OG08; Cod90].

- *Verletzung von Domänen Integritätsregeln*

  Eine Verletzung von Domänenintegritätsregeln liegt vor, wenn die Werte eines Attributes außerhalb eines vordefinierten Wertebereichs liegen [Cod90, Kapitel 3.1-3.2].

Bei der Evaluation der Konsistenz wird für jeden Wert überprüft, ob einer der obigen Widersprüche vorliegt. Die Konsistenz wird anschließend mit der Verhältnis basierten Metrik (Abschnitt 2.5) quantifiziert. Die Verwendung der Metrik erfüllt alle für die Unterstützung des Domänenexperten benötigten Anforderungen an Qualitätsmetriken (Abschnitt 3.3.2).

Von der Komponente verwendbare Parameter:

Für die Auswertung der Vollständigkeit lassen aus den obigen Widersprüchen, auf welche die Daten überprüft werden können, mehrere Parameter identifizieren, welche vom IT-Experten und dem Domänenexperten gestellt werden können.

- *[Unterschiedliche Formate] Format (IT-Experte, Domänenexperte)*

  Der IT-Experte kann seine technische Expertise nutzen und ein Format für ein Attribut in Form von Regulären Ausdrücken stellen.

  Da der Domänenexperte über keine technische Expertise verfügt, muss die Angabe des Formates für ihn vereinfacht werden. Weniger komplexe Formate können jedoch auch von ihm (beispielsweise mithilfe von Wildcards) angegeben werden.

- *[Semantische Widersprüche] Regeln (Domänenexperte)*

  Für die Auswertung semantischer Widersprüche kann der Domänenexperte analog zu den Regeln der Korrektheit die DM Notation zur Angabe von Regeln verwenden.

- *[Verletzung von Entitäts Regeln] Primärschlüssel (IT-Experte)*

  Die Angabe der Primärschlüssel sollte vom IT-Experten übernommen werden.

- *[Verletzung von Entitäts Regeln] Eigenschaften von Primärschlüsseln (Domänenexperte)*

  Dem Domänenexperten wird die Option eingeräumt die Eigenschaften von Primärschlüsseln anzugeben, da für die Angabe ob Daten Duplikate oder NULL-Werte enthalten dürfen, keine technische Expertise benötigt wird.

- *[Verletzung von Referentiellen Integritätsregeln] Fremdschlüssel (IT-Experte)*

  Fremdschlüssel sollten vom IT-Experten angegeben werden. Dieser kennt sich am besten mit dem Konzept von Fremdschlüsseln aus und weiß, auf welchen Datensatz diese verweisen.

- *[Verletzung von Domänen Integritätsregeln] Wertebereich (Domänenexperte)*

  Der Wertebereich sollte vom Domänenexperte angegeben werden, da er von sich am besten mit der Domäne, aus welcher die Daten stammen auskennt. Der Domänenexperte kann beispielsweise seine Expertise verwenden, um für die Daten aus Tabelle 3.1 einen Wertebereich für das Attribut RPM zu definieren, da er sich mit den Maschinen auskennt und weiß, welche Werte realistisch erreicht werden können.

Korrekturoptionen für identifizierte Probleme:

Bei der Verletzung einer Regel kann eine Korrektur analog zu der Korrektur einer verletzten Regel der semantischen Konsistenz ausgeführt werden. Bei unterschiedlichen Formaten kann der Domänenexperte bei der Formatierung der Werte unterstützt werden.

Für die Behebung der weiteren Widersprüche wird ausschließlich die Löschung der Werte, oder die Modifikation der gestellten Parameter angeboten.

Aktualität

Die Aktualität beschreibt, inwiefern Daten sich durch zeitliche Unterschiede von den Entitäten, welche sie modellieren, verändert haben, und inwiefern sie aus zeitlichen Gründen für den aktuellen Anwendungsfall geeignet sind (Abschnitt 2.4.1).

Bei der Auswahl einer Metrik zur Auswertung der Aktualitätsdimension kann zwischen zwei Anwendungsfällen unterschieden werden [Eve05; HK11; BWPT98]:

1. *Mit festem Verfallsdatum*

   Die Daten sind nur innerhalb eines gewissen Zeitraumes für den aktuellen Anwendungsfall interessant. Die Aktualität der Daten muss daher nur innerhalb dieses Zeitraums modelliert werden [Eve05; HK11; BWPT98]. Bei Daten, die älter als dieser Zeitraum sind, wird davon ausgegangen, dass diese keinen Mehrwert liefern [Eve05; HK11; BWPT98].

   Beispielsweise sind für die Berechnung des monatlichen Umsatzes eines Unternehmens ausschließlich die Daten der letzten 30 Tage relevant.

2. *Ohne festes Verfallsdatum*

   Es existiert kein fester Zeitraum, für welchen die Daten gültig sind [Eve05; HK11]. Stattdessen kann davon ausgegangen werden, dass mit jeder vergangenen Zeiteinheit ein Teil der Daten ungültig (sich in der Real-Welt verändert) wird [Eve05; HK11].

   Beispielsweise kann bei Kundendaten eines Unternehmens kein Zeitraum für die Gültigkeit der Daten festgelegt werden. Es kann jedoch davon ausgegangen werden, dass alte Kundendaten sich über die Zeit mit einer höheren Wahrscheinlichkeit in der Real-Welt verändert haben (beispielsweise durch einen Umzug, eine neue Bankleitzahl oder neue Kontonummer) als neu erhobene Werte.

Es wird daher der hybride Ansatz (aus Abschnitt 2.5) von Even und Shankaranarayanan, die für beide Anwendungsfälle eine separate Metrik bereitstellen ausgewählt, um beide Anwendungsfälle unterstützen zu können.

Von der Komponente verwendbare Parameter:

Für die Auswertung der Aktualität lassen sich folgende Parameter, die durch den Domänen- und IT-Experten bereitgestellt werden können, identifizieren:

- *[Beide Ansätze] Alter der Daten (Domänenexperte, IT-Experte)*

  Für die Auswertung der Dimension wird das Alter der Daten benötigt. Wenn das Alter der Daten nicht direkt von der Datenquelle bereitgestellt wird, kann ein Attribut der Daten verwendet werden, welches das Alter der Daten enthält (falls ein solches existiert). Falls ein solches Attribut verwendet werden soll, muss dies durch den Domänenexperten oder den IT-Experten spezifiziert werden.

- *[Festes Verfallsdatum] Zeitraum (Domänenexperte)*

  Gibt es eine feste Dauer nach der Daten ungültig werden hängt diese entweder vom Anwendungsfall des Domänenexperten ab oder von Eigenschaften, welche aus der Domäne der Daten hervorgehen. In beiden Fällen ist der Domänenexperte am besten für die Bereitstellung des Parameters geeignet.

- *[Festes Verfallsdatum] Abnahme Geschwindigkeit (Domänenexperte)*

  Für die Modellierung des Verfalls der Daten innerhalb eines festen Zeitraums werden Informationen über die Geschwindigkeit, mit welcher die Qualität der Daten innerhalb dieses Zeitraums abnimmt, benötigt [Eve05; BWPT98]. Da diese von dem Kontext, in welchem die Daten verwendet werden, abhängig ist [Eve05], muss die Abnahme Geschwindigkeit vom Domänenexperten gestellt werden.

- *[Ohne festen Gültigkeitszeitraum] Abnahme Geschwindigkeit (Domänenexperte)*

  Ohne die Verwendung eines festen Gültigkeitszeitraums wird die probabilistische Metrik verwendet [Eve05]. Die Abnahme Geschwindigkeit kann in diesem Fall vom Domänenexperten als prozentualer Anteil der Werte pro Zeiteinheit (beispielsweise Tag, Monat, Jahr), welche ihre Gültigkeit verlieren, ausgedrückt werden [HK11]. Auch hier wird Domänenwissen benötigt um einschätzen zu können, wie schnell die Daten ihren Wert verlieren. Der Domänenexperte eignet sich daher am besten für die Angabe dieses Parameters.

Korrekturoptionen für identifizierte Probleme:

Da das Alter der Daten nicht verändert werden kann, ist die einzige Maßnahme zur Verbesserung der Aktualitätsdimension die Löschung alter Werte. Der Domänenexperte kann hierfür einen Zeitpunkt auswählen, ab welchem die Werte gelöscht werden sollen.

3.3.4 Enthaltene Komponenten (statistisch)

Die statistischen Komponenten beschreiben die Daten. Der Domänenexperte erhält so einen Überblick über die Eigenschaften der Daten und kann so ein Verständnis für die Daten entwickeln. In Kombination mit seiner Expertise über die Domäne, aus welcher die Daten stammen, kann der Domänenexperte mithilfe der beschriebenen Eigenschaften der Daten Probleme in den Daten identifizieren. Die Identifizierung von Problemen der Daten erfolgt durch den Abgleich der Eigenschaften der Daten mit den vom Domänenexperten auf Basis seiner Expertise erwarteten Eigenschaften. Weichen die von den statistischen Komponenten beschriebenen Eigenschaften der Daten maßgeblich von der Erwartung des Domänenexperten ab, liegt ein Problem in den Daten vor.

Für die Beschreibung der Daten werden Metriken der deskriptiven Statistik verwendet [Nic07]. Die Anwendung dieser Metriken findet ausschließlich auf der Attributs-Ebene der Daten statt. Bei der Anwendung der Metriken muss zwischen kategorischen Daten (beispielsweise Herkunftsland, Typ,…) und numerischen Daten (beispielsweise Alter, Größe,…) unterschieden werden [Nic07].

Für beide Datentypen kann die Häufigkeitsverteilung der Werte beschrieben werden und als Komponente in das Datenqualitäts-Repository integriert werden [Nic07]. Das Ergebnis der Auswertung dieser Komponente ist eine grafische Darstellung der Werteverteilung. Diese kann für kategorische Daten in Form eines Balkendiagramms und für nummerische Daten in Form eines Histogramms erfolgen [Nic07].

Bei nummerischen Daten können zusätzlich die Lage und die Streuung der Daten gemessen werden [Nic07].

Die Lage der Daten kann mithilfe mehrerer Maße gemessen werden [Nic07]. Folgende Metriken sind zu diesem Zweck als Komponente enthalten, um die Lage der Daten umfangreich zu beschreiben [Nic07]:

- arithmetisches Mittel
- Median
- Modus (häufigster Wert)
- Quartile

Zur Messung der Streuung der Daten stehen ebenfalls mehrere Maße zur Verfügung [Nic07]:

- Varianz
- Standardabweichung
- Abstand zwischen Minimum und Maximum
- Interquartilsabstand

Zusätzlich ist eine grafische Darstellung der Lage und Streuung der Daten in Form eines Box-Plots möglich [Nic07]. In diesem sind Informationen zum Median, den Quartilen, dem Interquartilsabstand sowie zu Ausreißern enthalten [Nic07].

3.4 Benutzeroberfläche

Die Benutzeroberfläche (Abbildung 3.1 K3) ist ein kritischer Bestandteil für die Erfüllung der Anforderungen des Domänenexperten (Abschnitt 1.1). Die Erfüllung der Anforderung A4 „verständliche Darstellung und Bedienung“ hängt ausschließlich von der Implementierung der Benutzeroberfläche ab. Die Erfüllung weitere Anforderungen (A1, A2, A6, A7) ist zumindest teilweise von der Implementierung der Benutzeroberfläche abhängig.

3.4.1 Aufgabe

Die Benutzeroberfläche muss die im Prozessmodell (Abschnitt 3.1) dargestellten Interaktionen des Domänen- und IT-Experten durch Bereitstellung der entsprechenden (grafischen) Dialoge unterstützen. Anschließend muss die Benutzeroberfläche das resultierende Feedback zur Datenqualität grafisch in den Analyseprozess des Domänenexperten integrieren. Um diese Aufgaben erfüllen zu können, muss die Benutzeroberfläche folgende drei Bestandteile beinhalten:

- Dialoge zur Bearbeitung von Metadaten (Domänenexperte)
- Dialoge zur Bearbeitung von Metadaten (IT-Experte)
- Integration des Datenqualitäts-Feedbacks in den Workflow des Domänenexperten

Im Folgenden wird eine den Anforderungen (Abschnitt 1.1) des Domänenexperten konforme Implementierung dieser Bestandteile genau definiert.

3.4.2 Dialoge zur Bearbeitung von Metadaten (Domänenexperte)

Damit der Domänenexperte die für die Anwendung der Komponenten des Datenqualitäts-Repositorys (Abschnitt 3.3) notwendigen Metadaten bereitstellen kann, müssen die entsprechenden Dialoge für die Angabe der einzelnen Parameter in der Benutzeroberfläche existieren.

Beim Design dieser Dialoge muss vor allem auf eine einfache Verständlichkeit (Abschnitt 1.1, A4) dieser geachtet werden. Während der IT-Experte Metadaten mithilfe technischer Ressourcen, wie beispielsweise regulären Ausdrücken, Metadaten angeben kann, müssen die Dialoge des Domänenexperten so weit wie möglich vereinfacht werden, um (möglichst) keine technische Expertise vorauszusetzen. Die Eingabe von Parametern, welche nicht ohne technische Expertise vollzogen werden kann, sollte dem IT-Experten überlassen werden, um Verwirrungen und potenzielle Fehler zu vermeiden. Die Dialoge zur Eingabe solcher Parameter sollten dem Domänenexperten daher nicht, oder nur in eingeschränkter Form, zur Verfügung stehen (etwa durch die Angabe eines Formates mithilfe von Text und vorgefertigten Wildcards anstatt von Regulären Ausdrücken).

Angelehnt an die Funktionalität von Tableau Desktop[4] können für häufig verwendete Attribute der Daten (Stadt, Land, Postleitzahl,...) zur Angabe der Metadaten dieser Attribute Vorlagen genutzt werden, welche die für die Auswertung relevanten Metadaten (Referenzdaten, Formate, Wertebereich,...) enthalten. Der Domänenexperte muss für Attribute, welche als Vorlage enthalten sind, nur die für das Attribut der Daten passende Vorlage auswählen (Abbildung 3.3), um die Auswertung der Datenqualität zu ermöglichen. Der IT-Experte kann weitere Vorlagen für innerhalb der von ihm verwalteten Datenbanken häufig auftretende Attribute definieren (beispielsweise „Maschinentyp“ oder Firmeninterne IDs).

Um dem Domänenexperten eine verständliche Verwaltung der Metadaten zu ermöglichen, müssen nicht nur die Dialoge zur Anpassung der Metadaten nachvollziehbar sein, sondern auch die Benutzeroberfläche, in welche diese Dialoge integriert werden (Abschnitt 1.1, A4). Für eine einfache Bedienung dieser ist wichtig, dass der Domänenexperte aktiv auf Attribute, für welchen Metadaten

[4] https://help.tableau.com/current/prep/en-us/prep_validate_data.htm

Abbildung 3.3: Exemplarischer Dialog für die Auswahl einer Vorlage durch den Domänenexperten mit der vom IT-Experten definierten Vorlage „Maschinentyp“

angegeben werden können, hingewiesen wird. Nicht benötigte oder mit anderen Metadaten in Konflikt stehende Dialoge sollten ausgeblendet werden (beispielsweise sollte der Domänenexperte keinen numerischen Wertebereich für ein Text-basiertes Attribut festlegen können).

Die Abbildung 3.4 stellt eine prototypische Benutzeroberfläche zur Angabe von Metadaten dar. Dieses enthält bereits eine Reihe von exemplarischen Dialogen zur Angabe von Metadaten. Im Folgenden wird die ein Konzept für die Benutzeroberfläche genauer vorgestellt.

Das Konzept für die Angabe von Metadaten (Abbildung 3.4) zeigt dem Domänenexperten eine tabellarische Übersicht über die Datenquelle. Die Übersicht beinhaltet die Attribute und eine Teilmenge der Attributwerte, für welche die Metadaten bereitgestellt werden (1). Der Domänenexperte wird innerhalb der Übersicht auf Attribute hingewiesen (rot), für welche er weitere Metadaten angeben kann.

Zur Eingabe der Metadaten (2) kann der Domänenexperte das Attribut in der Übersicht (1) anklicken oder im Dropdown Menü (B) auswählen. Der Domänenexperte erhält daraufhin Dialoge, die ihm die Angabe der Metadaten für das Attribut ermöglichen. Das Konzept enthält unter anderem einen Dialog zur Verwendung von Vorlagen für Attribute (C) und die einfache Erstellung eines Formates für die Attributwerte (D, exemplarisch eine Kombination aus zwei Zahlen). Ihm wird außerdem angezeigt, dass der IT-Experte bereits Referenzdaten für das ausgewählte Attribut („column 2“) definiert hat (E). Falls sich die Metadaten des IT-Experten später als inkorrekt herausstellen sollten, erhält der Domänenexperte die Option den IT-Experten zur Überprüfung dieser aufzufordern oder sie für die weitere Auswertung der Datenqualität zu ignorieren (Abschnitt 2.6, Prinzip 1).

Der Domänenexperte hat die Option, Regeln für das Attribut zu definieren (F). Die Angabe der Regel erfolgt mithilfe von „DMN decision tables“. Das Konzept stellt exemplarisch die Regeln column 1=„example val“ –> column 2<=50 und column 2=„example val2“ –> column 2=Null dar.

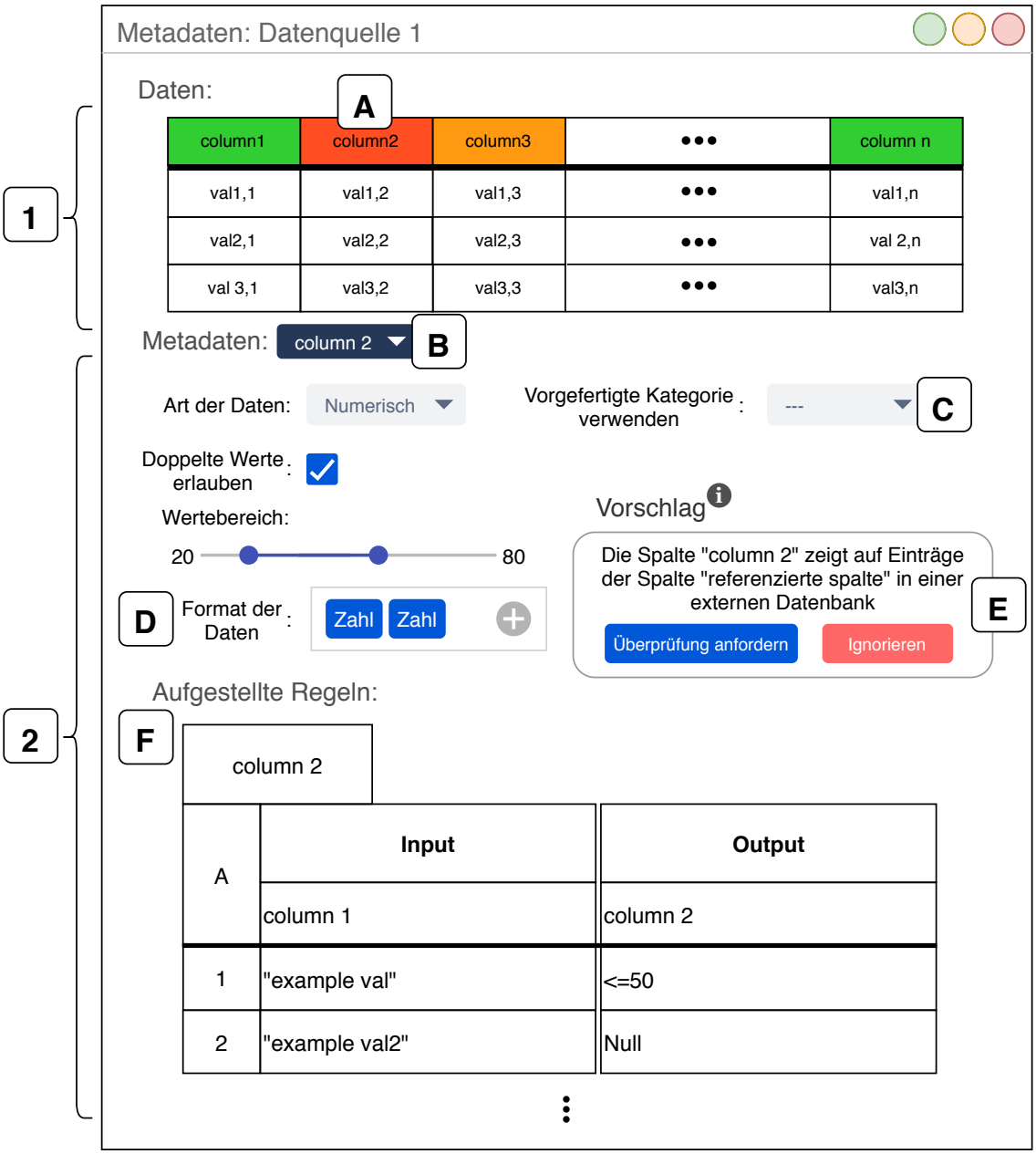

Abbildung 3.4: Konzept für eine Benutzeroberfläche zur Angabe von Metadaten mit Übersicht über die Daten (1) und Dialogen zur Angabe der Metadaten durch den Domänenexperten (2). Der Prototyp beinhaltet verschiedene Dialoge zur Angabe von Metadaten (C, D), die Anzeige der komplexer Metadaten des Domänenexperten (E) und die Erstellung benutzerdefinierter Regeln (F).

3.4.3 Dialoge zur Bearbeitung von Metadaten (IT-Experte)

Die Dialoge zur Bearbeitung von Metadaten des IT-Experten werden im Gegensatz zu den Domänenexperten nicht durch die technischen Kenntnisse des Anwenders limitiert. Die Metadaten können vom IT-Experte daher ohne die Verwendung vereinfachter Dialoge angegeben werden. Für den Domänenexperten ist daher eine Bearbeitungsoption der Metadaten über eine Konsolen-Schnittstelle oder die Bearbeitung der Metadaten im XML Format zur Angabe von Metadaten in der Benutzeroberfläche ausreichend. Das Hinzufügen neuer Datenquellen kann vom IT-Experten ebenfalls durch eine Konsolen-Schnittstelle bewerkstelligt werden. Optional können für den IT-Experten jedoch auch komplexere Dialoge integriert werden, welche in der Lage sind komplexere Parameter abzufragen.

der IT-Experten die verfügbaren Metadaten daher ohne die Verwendung vereinfachter Dialoge

3.4.4 Integration des Datenqualitäts-Feedbacks in den Workflow des Domänenexperten

Die Integration eines Feedbacks über die Datenqualität in den vom Domänenexperten modellierten Workflow soll den Domänenexperten bei der Verarbeitung und Auswertung der Daten unterstützen. Die Unterstützung erfolgt durch die Aufklärung des Domänenexperten über potenzielle Datenqualitätsprobleme sowie eine Rückmeldung zu den Auswirkungen verschiedener Arbeitsschritte. Auf Basis der Anforderungen des Domänenexperten (Abschnitt 1.1) lassen sich mehrere Kriterien für die Integration eines Feedbacks über die Datenqualität in den Analyseprozess identifizieren:

K1: Integration in den vollständigen Analyseprozess

K2: Feedback auf verschiedenen Granularitätsebenen

K3: Verständliche Darstellung und verständliche Messgrößen

K4: Korrekturoptionen

K5: (Optional) Zeitnahe Rückmeldung

Im Nachfolgenden wird die Notwendigkeit und Umsetzung der oben aufgeführten Kriterien erläutert. Die Umsetzung der Kriterien wird anhand von grafischen Konzepten einer Konformen Implementierung veranschaulicht.

K1 Integration in den vollständigen Analyseprozess
Aus der Anforderung A1 (Abschnitt 1.1) des Domänenexperten geht die Notwendigkeit hervor, den Domänenexperten während des kompletten Analyseprozesses über die Datenqualität zu Informieren. Zusätzlich muss der Domänenexperte die Auswirkungen angewandter Operationen auf die Datenqualität beurteilen können. Für die Erfüllung des Kriteriums K1 wird daher die Darstellung der Datenqualität vor und nach jeder angewandten Transformation gefordert. Nur auf diese Weise kann der Domänenexperte die Auswirkung der Operation auf die Datenqualität erkennen.

Einen Konzept für eine Benutzeroberfläche, die dieses Kriterium erfüllt, stellt Abbildung 3.5 dar. Der Domänenexperte erhält in diesem Prototyp anhand der UI-Elemente A, C eine Rückmeldung über die Qualität der Daten und kann so die Auswirkung seiner Filter-Operation (B) auf die Datenqualität genau einschätzen (in Abbildung 3.5 exemplarisch eine deutliche Verbesserung).

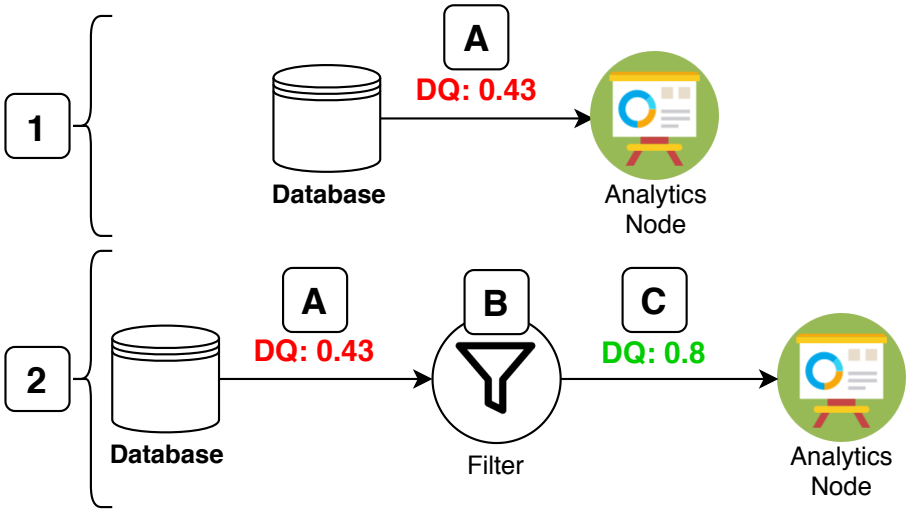

Abbildung 3.5: Zwei einfache Workflows (1, 2) zur Analyse einer Datenbank mit Feedback zur Datenqualität (A, C) sowie jeweils einem Filterknoten (B)

K2 Feedback auf verschiedenen Granularitätsebenen
Das Kriterium K2 (basierend auf A2 Abschnitt 1.1) setzt die Integration des Feedbacks auf verschiedenen Granularitätsebenen voraus. Die Benutzeroberfläche muss die Anzeige der Datenqualität auf den verschiedenen Granularitätebenen unterstützen. Im Folgenden wird daher ein Konzept zur Integration eines Feedbacks über die Datenqualität auf verschiedenen Granularitätebenen vorgestellt. Das Kriterium K1 wird von dem Prototyp weiterhin erfüllt.

Die Granularität, mit welcher das Feedback zur Datenqualität angezeigt wird, richtet sich bei diesem Konzept nach den Bedürfnissen des Domänenexperten. Die Datenqualität kann vom Domänenexperten jederzeit explorativ auf den verschiedenen Ebenen betrachtet (unterstützt die Exploration der Daten nach Abschnitt 2.6, „Explorative Character“) werden. Die explorative Betrachtung kann eine einfachere Identifizierung von Problemstellen ermöglichen. Insgesamt stehen dem Domänenexperten die vier folgenden Granularitätsebenen zur Verfügung:

Workflow Ebene

Die Workflow Ebene stellt die gröbste Granularitätsebene dar. Sie betrachtet den kompletten, vom Domänenexperten modellierten Prozess (den Workflow).

Das Feedback zur Datenqualität ist auf dieser Ebene auf einen einzelnen Wert aggregiert (Abbildung 3.5 A, C). Dies ist ausreichend, um dem Domänenexperten einen Überblick über die Datenqualität in jedem Schritt des Workflows zu verschaffen und ihn auf Problemstellen (beispielsweise Abbildung 3.5 A) aufmerksam zu machen. Der Domänenexperte erhält ein ausreichendes Feedback, um Auswirkung verschiedener Komponenten auf die Datenqualität einfach einschätzen zu können. Gleichzeitig lenkt die Datenqualität ihn jedoch nur minimal vom Analyseprozess ab.

Ein Beispielszenario stellt Abbildung 3.5 dar. Hier ist der Domänenexperte im Workflow 1 Punkt A auf die schlechte Datenqualität aufmerksam gemacht worden. Er reagiert daher in Workflow 2 mit einer Filter-Operation (Abschnitt 3.5, B) auf die schlechte Qualität der Daten

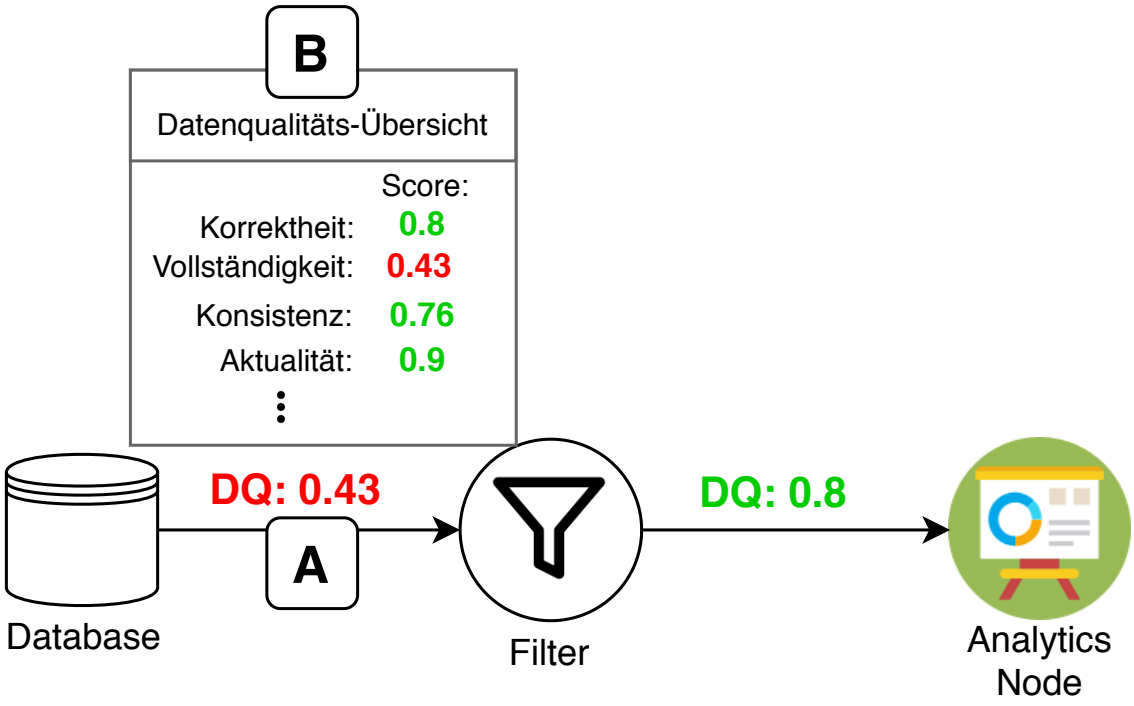

Abbildung 3.6: Ein Workflow mit aggregiertem Feedback (A) und einer Übersicht der verschiedenen Dimensionen (B) aus welchen sich A zusammensetzt

aus Abbildung 3.5 Workflow 1 um fehlerhafte Daten zu entfernen und so die Datenqualität zu verbessern. Er bekommt in Abbildung 3.5 C ein direktes Feedback zum Erfolg seiner Operation.

Dimensionsebene
Der Domänenexperte kann das Feedback zur Datenqualität aus der Workflowebene verfeinern, in dem er die Datenqualitäts-Werte der einzelnen Dimensionen (Abbildung 3.6, B) betrachtet. Diese bilden die Grundlage für den aggregierten Wert der Workflow Ebene (Abbildung 3.6, A).

Die Darstellung einzelner Qualitäts-Dimensionen bietet dem Domänenexperten einen besseren Überblick über die Datenqualität. Es lassen sich zusätzlich Gründe für Datenqualitätsprobleme genauer identifizieren. So lässt sich in Abbildung 3.6 (B) beispielsweise die Vollständigkeit der Daten als Grund der schlechten aggregierten Datenqualität (A) ausfindig machen. Für die Aggregation der Dimensions-Werte wird in Abbildung 3.5 exemplarisch das Minimum verwendet.

Attributsebene
Die Betrachtung der Datenqualität auf der Attributsebene verfeinert die Darstellung der Dimensionsebene weiter, indem Datenqualitätsmerkmale für jedes Attribut der Datenquelle dargestellt werden.

Der Prototyp Abbildung 3.7 stellt die wichtigsten Merkmale dieser Ebene dar. Dieser beinhaltet für jedes Attribut (A) jeweils die Auswertungen der Datenqualitätsdimensionen auf der Attributsebene sowie Funktionen der deskriptiven Statistik (C). Die Auswertung der

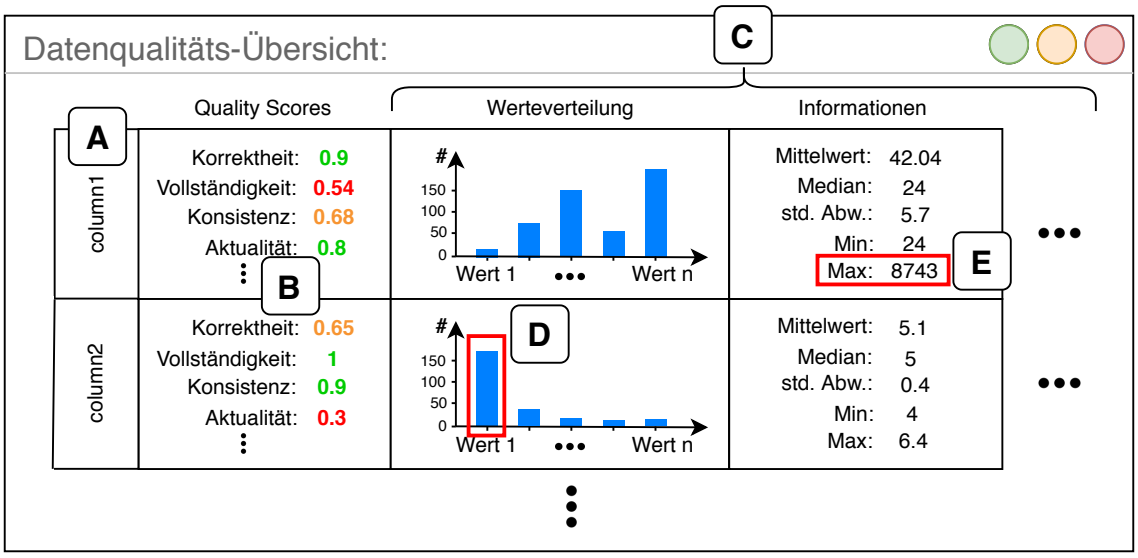

Abbildung 3.7: Ein Datenqualitätsübersichtsdialog für Attribute (A) bestehend aus den quantifizierten Werten der Qualitätsdimensionen (B), Funktionen der deskriptiven Statistik (C) sowie Werten, die der Domänenexperte als inkorrekt identifiziert (D, E)

Datenqualitätsdimensionen auf der Attributsebene (B) ermöglicht es dem Domänenexperten Datenqualitätsprobleme auf einzelne Attribute zurückzuführen. Die Funktionen der deskriptiven Statistik (C) beschreiben Eigenschaften des jeweiligen Attributes. Kombiniert mit der Expertise des Domänenexperten können durch diese Funktionen inkorrekte Werte identifiziert werden.

Ein Beispiel für Werte, die der Domänenexperte als inkorrekt identifizieren kann, ist in Abbildung 3.7 in D und E gegeben. Der Domänenexperte weiß in diesem Beispiel aufgrund seiner Expertise, dass das Maximum für „column1“ (E) deutlich zu hoch ist. Er erkennt außerdem, dass der Wert „Wert1“ in „column2“ (D) häufiger vorkommt als er eigentlich sollte.

Datenebene

Der Analyseprozess des Domänenexperten kann an mehreren Stellen eine direkte Interaktion mit (einer Teilmenge) der Daten in tabellarischer Form (beispielsweise zur interaktiven Bearbeitung) beinhalten. Das Feedback über die Datenqualität muss daher auch für die direkte Darstellung (einer Teilmenge) der Daten in tabellarischer Form in die Benutzeroberfläche integriert sein.

Um auf dieser Ebene ein sinnvolles Feedback zur Datenqualität geben zu können, wird die Datenqualität nicht nur auf der Attributebene (Abbildung 3.8, A), sondern auch für einzelne Tupel (Abbildung 3.8, B) berechnet. Dies ermöglicht eine genauere Identifizierung von Problemstellen. Die Datenebene ermöglicht die Implementierung von zusätzlichen Funktionen, um die Exploration der Datenqualität zu unterstützen. Der Domänenexperte kann beispielsweise die Tupel nach ihrer Datenqualität sortieren oder gezielt Tupel mit schlechter Qualität anzeigen lassen. Für die Datenqualität auf der Attributsebene (A) und der Tupelebene (B) kann der Domänenexperte Übersichtfenster analog zu Abbildung 3.6, (B) aufrufen.

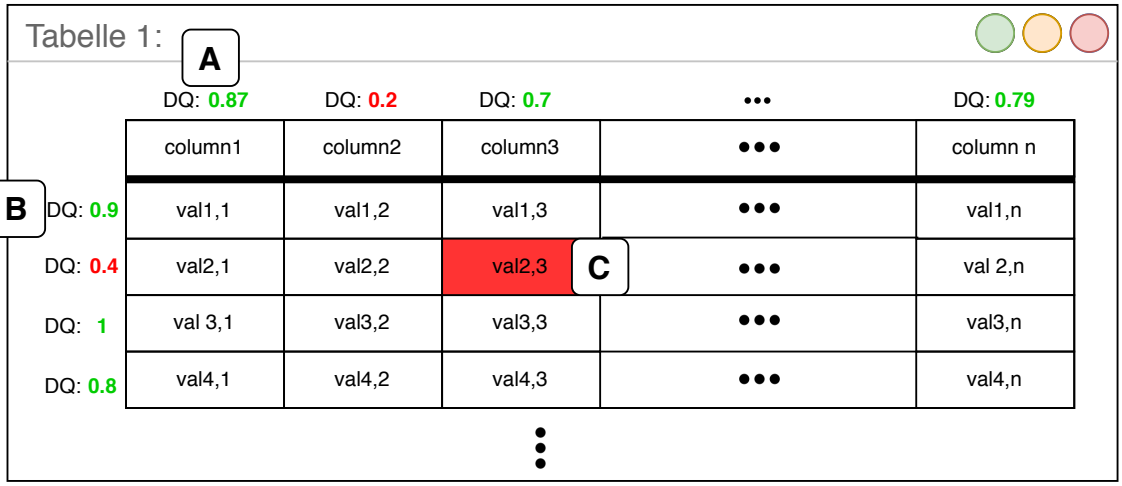

Abbildung 3.8: Ein Dialog zur direkten Interaktion mit den Daten (beispielsweise zum Aufbereiten von Daten) mit einem Feedback zur Datenqualität auf Attributsebene (A), auf Tupelebene (B) sowie auf der Attributwertebene (C)

Der Prototyp Abbildung 3.8 stellt in C exemplarisch die Darstellung von gefundenen Fehlern für einzelne Attributwerte (in C eine verletzte Abhängigkeit) dar. Der Domänenexperte kann mit diesen interagieren um Informationen zur Art der gefundenen Problemstelle einzusehen.

K3. Verständliche Darstellung und verständliche Messgrößen

Der Domänenexperte besitzt keine Erfahrung im Umgang mit Datenqualitätsproblemen. Die verständliche Darstellung und Bedienung wurde in Abschnitt 1.1, A4 daher als eine Anforderung an die Integration der Datenqualitätsüberwachung identifiziert. Die Erfüllung dieser Anforderung wird maßgeblich von der Implementierung der Benutzeroberfläche bestimmt, da diese die mit dem Domänenexperten interagierenden Elemente der Datenqualitätsüberwachung enthält. Bei der Erstellung von Dialogen für die Benutzeroberfläche muss daher auf eine verständliche Darstellung von Informationen und Interaktionsmöglichkeiten geachtet werden. Es sollten außerdem keine technisch komplexen Messgrößen für die Beschreibung von Eigenschaften der Daten verwendet werden.

K4. Korrekturoptionen
Um die Anforderungen des Domänenexperten zu erfüllen, müssen gefundene Datenqualitätsprobleme behoben werden können (Abschnitt 1.1, A7). Die Benutzeroberfläche muss deshalb die Interaktion des Domänenexperten mit den in der Evaluationsumgebung (Abbildung 3.1, K5) generierten Korrekturvorschlägen ermöglichen. Um diese Interaktion gewährleisten zu können, muss ein Dialog zur Korrektur von Datenqualitätsproblemen für den Domänenexperten verständlich sein (Abschnitt 1.1, A4). Der verwendete Aufbau der Dialoge beinhaltet daher immer einen Teil der die Problemstelle (möglichst einfach) beschreibt, gefolgt von Lösungsvorschlägen. Im Nachfolgenden wird der Aufbau der Dialoge dargestellt.

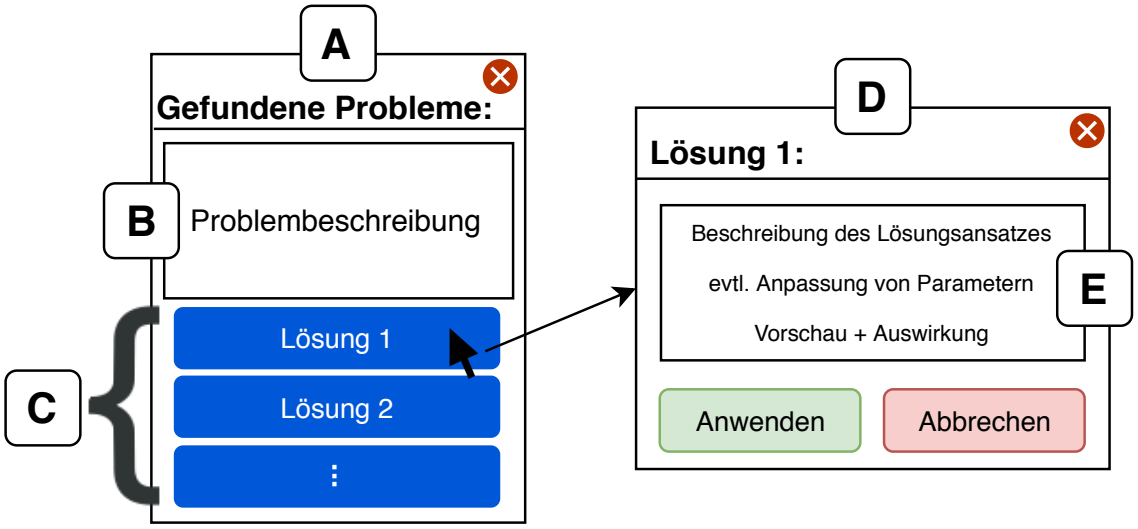

Abbildung 3.9: Konzept eines Dialogs für Korrekturoptionen der Darstellung des Problems A mit einer Beschreibung (B) und den generierten Lösungsvorschlägen (C). Der Dialog D stellt die Ausführung einer Korrekturoption mit den entsprechenden Informationen (E) dar

Aufbau von Korrektur-Dialogen

Zur Korrektur von Datenqualitätsproblemen wird ein Aufbau bestehend aus zwei Dialogen vorgeschlagen (Abbildung 3.9, A, D). Der erste Dialog erklärt dem Domänenexperten das gefundene Datenqualitätsproblem (B) und listet die gefundenen Lösungsvorschläge (C) auf. Der Domänenexperte soll in diesem Dialog ein Verständnis für das gefundene Problem und mögliche Korrekturvorschläge bekommen. Die Korrekturvorschläge können bereits in diesem Dialog durch Tooltips kurz erklärt werden. Der Zweite Dialog (D) öffnet sich bei der Auswahl einer Korrekturoption durch den Domänenexperten. Er erklärt dem Domänenexperten, was bei Anwendung der Korrektur passiert und bietet ihm eine Vorschau zu Änderungen der Datenqualität und der Daten (E). Der zweite Dialog ist (zusätzlich zu der Anforderung A4 aus Abschnitt 1.1) wichtig für die Erfüllung des Prinzips „Put the User in Charge“ (Abschnitt 2.6), da der Domänenexperte nur dann (sinnvoll) die vollständige Kontrolle über die Anwendung von Korrekturmaßnahmen übernehmen kann, wenn ihm die Funktionsweise und die Auswirkungen der jeweiligen Korrekturoption bekannt ist.

Beispiel

Ein Beispiel für die Implementierung des Konzepts für die Verletzung einer Abhängigkeit stellt Abbildung 3.10 dar. Der Dialog A erklärt das Problem in kompakter Form. Eine genaue Erklärung des Problems bietet Tooltip (B). Dieser kann vom Domänenexperten nach Bedarf zu Hilfe herangezogen werden. Im Beispiel handelt es sich um die verletze Abhängigkeit des Attributs „column3“ von den Attributen „column1“ und „column2“. Der Wert aus „column3“ entspricht nicht dem aufgrund der Abhängigkeit erwartetem Wert „Wert3“ (sondern „notWert3“).

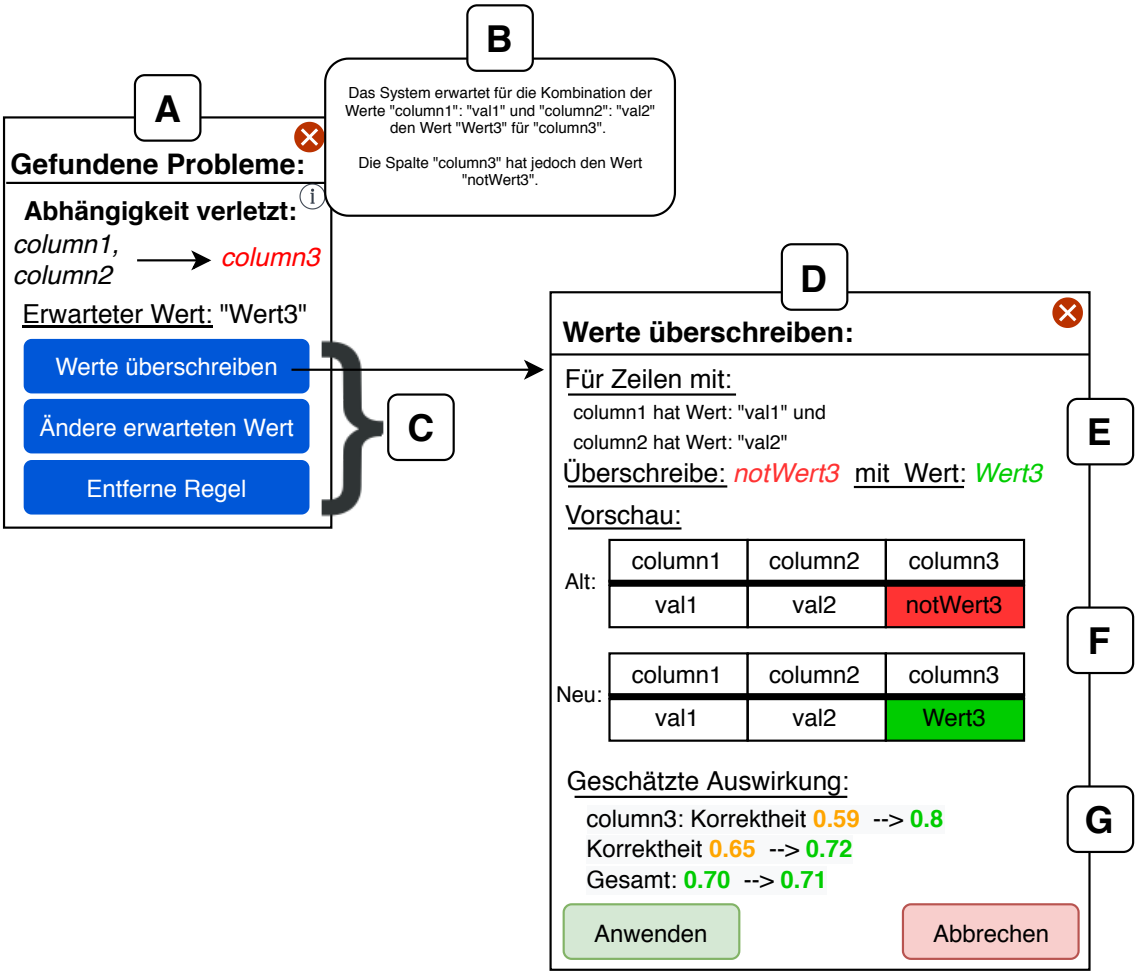

Abbildung 3.10: Beispiel für die Implementierung des Konzeptes aus Abbildung 3.9 anhand einer verletzten Abhängigkeit. Mit der Darstellung des Problems in A. Die Darstellung beinhaltet einen Tooltip (B) zur vereinfachten Erklärung des Problems und Korrekturoptionen (C). Der Dialog für die Anwendung einer Korrekturoption (exemplarisch „Werte überschreiben“, in D). Die Anwendung der Korrkturoption beinhaltet die Erklärung dieser (E), eine Vorschau (F) und eine Vorhersage zu den Auswirkungen der Anwendung (G).

Die Korrekturoptionen für diese Problem (C) beinhalten das automatische Überschrieben des Wertes „notWert3“ mit dem erwartetem Wert („Wert3“), die Änderung des für die spezifische Kombination der Werte der beiden Attribute erwarteten Wertes (zu „notWert3“) und das Entfernen der Abhängigkeit.

Für die Korrekturoption „Werte überschreiben“ stellt D den Dialog zur Anwendung zur Verfügung. Der Domänenexperte bekommt in diesem Dialog eine Erklärung, was bei der Ausführung der Korrekturoption passiert (E), zusammen mit der Vorschau einer Änderung (F) und geschätzten Auswirkungen auf die Qualität der Daten (G).

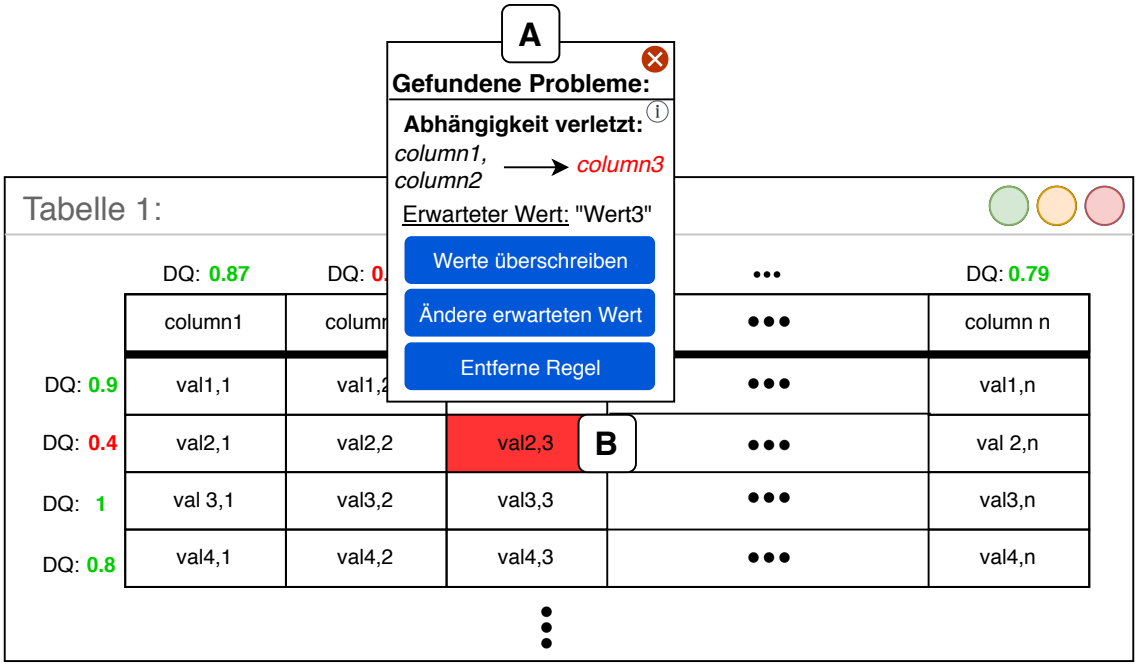

Abbildung 3.11: Integration von Korrekturdialogen auf Datenebene mit dem Dialog A, welcher Korrekturoptionen für eine spezifische Problemstelle (B) liefert

Die meisten Datenqualitätsprobleme beziehen sich auf einzelne Attributwerte. Die Integration von Korrekturoptionen in Granularitätsebenen, welche die Datenqualität mindestens auf der Ebene der einzelnen Attribute darstellen, ist deshalb am geeignetsten. Bei gröberen Granularitätsebenen ist der Einfluss von Korrekturoptionen auf die aggregierte Bewertung der Datenqualität schwerer ersichtlich, da diese sich aus der Datenqualität mehrerer Attribute zusammensetzen. Die Integration von Korrekturoptionen trägt außerdem bei den gröberen Granularitätebenen nicht zu der schnellen Beurteilung der Datenqualität (eigentliches Ziel dieser) bei. Die Korrekturoptionen werden daher nur auf der Attributs- und der Datenebene in die Benutzeroberfläche integriert.

Korrekturoptionen auf Datenebene Bei der Integration von Korrekturoptionen auf Datenebene kann die präzise Darstellung der Problemstellen mit der Darstellung von entsprechenden Korrekturoptionen verbunden werden. Der Domänenexperte kann durch Interaktion (beispielsweise Rechts-click, Hovern, ...) mit markierten Datenqualitätsproblemen einen Korrekturdialog öffnen. Die Integration auf Datenebene ist exemplarisch in Abbildung 3.8 dargestellt.

Korrekturoptionen auf Attributsebene Die Attributsebene bietet eine Übersicht über die Datenqualität für einzelne Attribute. Im Gegensatz zur Datenebene können daher keine konkreten Problemstellen hervorgehoben werden. Für die Integration der Korrekturoptionen in die Attributsebene ist es daher sinnvoll, dem Domänenexperten eine Übersicht über die gefundenen Korrektursoptionen (jeweils für jeder Attribut) zu bieten. Damit der Domänenexperte die Korrekturoptionen der Übersicht effektiv nutzen kann müssen die angezeigten Korrekturoptionen nach ihrem Einfluss auf die Datenqualität sortiert sein. Bei einer willkürlichen Reihenfolge der Korrekturoptionen können

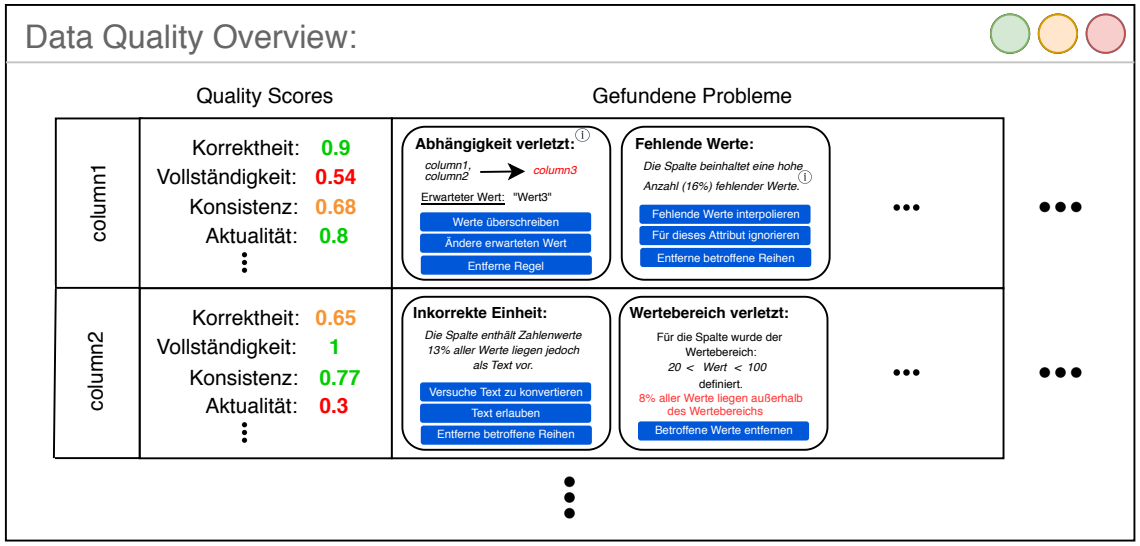

Abbildung 3.12: Integration von Korrekturoptionen auf Attributsebene mit nach Fehlerhäufigkeit sortierten Dialogen für die Attribute „column1“ und „column2“

schwerwiegende Fehler leicht übersehen werden. Die Sortierung der Korrekturoptionen nach ihrem Einfluss kann über die Fehlerhäufigkeit oder die maximal erzielbare Verbesserung der Datenqualität vorgenommen werden. Ein Beispiel für die Integration von Korrekturoptionen auf Attributsebene stellt Abbildung 3.12 dar.

K5. Zeitnahe Rückmeldung

Die Zufriedenheit von Nutzern einer Anwendung sinkt mit steigender Antwortdauer [HD00]. Des Weiteren senkt eine hohe Antwortzeit die Effizienz, da der Domänenexperte langsamer auf Probleme reagieren kann. Das Feedback zur aktuellen Datenqualität sollte deshalb schnellstmöglich auf Aktionen des Domänenexperten reagieren. Die Zufriedenstellung dieser Anforderung hängt nicht nur von einer zeitnahen Aktualisierung des angezeigten Feedbacks in der Benutzeroberfläche, sondern auch von der von der Evaluationsumgebung benötigten Zeit zur Erstellung des Feedbacks ab. Dies stellt vor allem unter dem „Volume “ Aspekt von „Big Data“ (Abschnitt 2.1) eine Herausforderung dar [TKS+16].

3.5 Datenqualitäts-Service

Der Datenqualitäts-Service stellt den korrekten Ablauf des Prozessmodells 3.1 sicher. Er dient als Schnittstelle zwischen der Benutzeroberfläche, dem Datenquellen-Repository, dem Datenqualitäts-Repository und der Evaluationsumgebung und koordiniert den Austausch von Informationen zwischen diesen.

3.5.1 Aufgabe

Der Datenqualitäts-Service erstellt aus den gesammelten Informationen über eine Datenquelle ein datenquellenspezifisches Auswertungsmodell. Diese beinhaltet die Komponenten, welche zur Auswertung der Datenqualität einer Datenquelle genutzt werden und die entsprechenden Parameter. Zur Identifikation der verwendeten Komponenten und Parameter wird das Datenqualitäts-Repository (Abschnitt 3.3) durchsucht.

Um den korrekten Ablauf des Prozesses sicherzustellen muss der Datenqualitäts-Service die Informationen dem Prozessmodell entsprechend weiterleiten, verarbeiten und anfordern. Dies erfordert die Unterstützung einer Reihe von Interaktionen zwischen den verschiedenen Komponenten. Nachfolgend wird ein Überblick über die Interaktionen, die vom Datenqualitäts-Service unterstützt werden müssen, gegeben:

Hinzufügen einer neuen Datenquelle

Beim Hinzufügen einer neuen Datenquelle durch den Domänenexperten (ausgehend von der Benutzeroberfläche) muss der Datenqualitäts-Service die zur neuen Datenquelle vorhandenen Metadaten aus dem Datenquellen-Repository abrufen.

Identifizierung anwendbarer Komponenten und benötigter Parameter

Nachdem eine neue Datenquelle hinzugefügt wurde, oder Informationen zu einer vorhandenen Datenquelle geändert wurden durchsucht der Datenqualitäts-Service das Datenqualitäts-Repository nach Komponenten, die zur Auswertung der Datenqualität genutzt werden können. Er identifiziert außerdem Parameter, welche der Domänenexperte zur Verbesserung der Auswertung spezifizieren kann.

Domänenexperte zur Spezifikation identifizierter Parameter auffordern

Für neue Datenquellen muss der Datenqualitäts-Service den Domänenexperten über die Benutzeroberfläche aktiv zur Spezifikation der Parameter auffordern (Abschnitt 1.1, A5).

Auswertungsmodell für Datenquelle übermitteln

Das vom Datenqualitäts-Service erzeugte Auswertungsmodell muss zur Auswertung der Datenqualität an die Evaluationsumgebung übermittelt werden.

Änderung eines Parameters

Ändert der Domänenexperte einen Parameter einer Datenquelle, muss der Datenqualitäts-Service das Auswertungsmodell entsprechend anpassen, und die Änderung an die Evaluationsumgebung übermitteln.

Metadaten speichern

Die während des Analyseprozesses vom Domänenexperten spezifizierten Parameter sollten im Datenquellen-Repository gespeichert werden, um eine erneute Eingabe der Parameter bei einer späteren Verwendung der Datenquelle zu vermeiden.

3.6 Evaluationsumgebung

Die Evaluationsumgebung wertet die Datenqualität im Workflow des Domänenexperten aus. Für die Evaluation werden die vom Datenqualitäts-Service generierten Auswertungsmodelle der einzelnen Datenquellen verwendet. Die Evaluation der Datenqualität im Workflow des Domänenexperten wird im Folgenden beschrieben:

Auswertung der Datenqualität im Workflow

Die Evaluation der Datenqualität muss für jede Datenquelle im Workflow und jeweils nach allen Operationen, welche im Workflow vorhandene Daten transformieren (die Datenqualität beeinflussen können) stattfinden. Dies wird benötigt um ein Feedback zur Datenqualität für den kompletten Workflow des Domänenexperten geben zu können (Abschnitt 1.1, A1). Für die Auswertung der Datenqualität an einer Stelle im Workflow können die im Auswertungsmodell enthaltenen Komponenten mit den enthaltenen Parametern ausgeführt werden, um Problemstellen, Korrekturoptionen und Auswertungsergebnisse zu erhalten. Diese Informationen können in die Benutzeroberfläche integriert werden, um dem Domänenexperten ein Feedback zur Datenqualität zu liefern.

Eine Herausforderung bei der Auswertung der Datenqualität im Workflow stellen Operationen, aus welchen Daten mit abgeänderten Attributen oder Eigenschaften resultieren. Ein Beispiel für ein solches Szenario ist die Kombination mehrerer Datenquellen mittels einer „Join“ Operation. Für die aus diesen Operationen entstehenden Daten ist kein Auswertungsmodell vorhanden.

Damit der Domänenexperte für solche Daten (möglichst) keine neuen Metadaten angeben muss, werden für solche Operationen die Metadaten und anwendbaren Komponenten aus den vorhandenen Datenquellen, welche die Grundlage der Daten bilden abgeleitet. Der Domänenexperte kann die abgeleiteten Metadaten jederzeit (wie bei anderen Datenquellen) abändern. Bei der Änderung der Metadaten einer Datenquelle aus welcher Metadaten abgeleitet werden, sollte die Änderung auch in den abgeleiteten Metadaten reflektiert werden (manuell vom Domänenexperten geänderte Parameter sollten dabei nicht ohne Rückfrage überschrieben werden).

Bei der Ableitung von Metadaten können mehrere Herausforderungen entstehen:

Kombinierte Attribut-Werte

Beinhalten die neuen Daten ein Attribut, dessen Wert sich aus mehreren verschiedenen Attributen alter Datenquellen zusammensetzt, kann dieses neue Eigenschaften besitzen. Eine Ableitung der Metadaten für solche Attribute ist daher nicht möglich. Der Domänenexperte muss die Parameter für diese Attribute daher manuell anlegen.

Ein Beispiel für dieses Verhalten ist die Konkatenation der Attribut-Werte „Tag“ und „Monat“ zum neuen Attribut „Datum“ der Form „Tag-Monat“. Ist beispielsweise für das Attribut „Monat“ der Wertebereich 1-12 definiert ist die Regel nicht auf das neue Attribut übertragbar.

Dasselbe gilt für neu erstellte Attribute, welche sich aus einer Teilmenge eines vorhandenen Attributes zusammensetzen. Beispielsweise „Monat“ aus „Datum“.

Konflikte zwischen Metadaten

Ein Attribut der neuen Daten kann in mehreren alten Datenquellen, die zur Erstellung der neuen Daten benutzt werden enthalten sein. In diesem Fall stehen mehrere Quellen zur Ableitung der Metadaten für das Attribut zur Verfügung. Enthalten diese Quellen unterschiedliche

Informationen über das Attribut ist eine automatische Ableitung der Metadaten für dieses Attribut nicht möglich. Die Auflösung eines solchen Konfliktes erfolgt am besten durch die Selektion der geeignetsten Quelle die Metadaten durch den Domänenexperten, da er sich mit den Daten auskennt und eine informierte Entscheidung auf Basis der Eigenschaften des neuen Attributes treffen kann.

Ein Beispiel für einen solchen Konflikt ist etwa die Vereinigung der Produktionsdaten zweier Fabriken, in welchen das Attribut „Datum“ vorhanden ist. Die Fabriken verwenden für das Attribtut „Datum“ jedoch verschiedene Formate (beispielsweise ”dd.mm.yyyyünd ”mm.dd.yyyy”). Da die Datenquellen zwei verschiedene Formate für das Attribut „Datum“ verwenden, unterscheiden sich auch die Metadaten für dieses Attribut zwischen den Datenquellen. Die Metadaten des Attributs „Datum“ können daher in den neuen Daten nicht eindeutig identifiziert werden.

Der Domänenexperte kann die Datenquelle, aus welcher die Metadaten abgeleitet werden sollen, angeben, oder die Metadaten für das neue Attribut selbst definieren.

Die Auswertung der Datenqualität kann für große Datenmengen mit einem erheblichen Zeit- und Rechenaufwand verbunden sein [TKS+16]. Dies ist vor allem im Kontext Big Data-Aspektes „Volume“ (Abschnitt 2.1) relevant. Es wird daher im Hinblick auf die großen Datenmengen im Kontext von Big Data ein Ansatz zur Auswertung der Datenqualität mithilfe von Stichproben der jeweiligen Datenquelle vorgeschlagen [TKS+16]. Die Auswertung der Datenqualität mithilfe von Stichproben wird im Folgenden näher erläutert:

Auswertung mithilfe von Stichproben

Das Ziel der Auswertung der Datenqualität auf Stichproben der Datenquelle ist die Reduzierung der zur Auswertung der Datenqualität benötigten Zeit und Rechenleistung [TKS+16]. Taleb et al. [TKS+16] stellen ein Prozessmodell zur Auswertung der Datenqualität mithilfe von Stichproben vor. Dieses wurde bereits im Abschnitt 2.8.3 erläutert. Im nachfolgenden wird eine adaptierte Version dieses Prozessmodells vorgestellt, welches in die hier vorgestellte Evaluationsumgebung integriert werden kann.

Das Prozessmodell (Abbildung 3.13) stellt den Auswertungsprozess (Abbildung 3.13, 2) der Evaluationsumgebung für die Datenqualität im Workflow (Abbildung 3.13, 1) dar. Zur Veranschaulichung wird die Berechnung der Datenqualität für einen einfachen zwei-elementigen Analyseworkflow, bestehend aus einer Datenquelle (1, A) und einem Auswertungsknoten (1, B), dargestellt. Der Workflow enthält daher nur eine Stelle (1, C), an der die Datenqualität berechnet werden muss. Für komplexere Workflows wird der Auswertungsprozess (2) jeweils an den Stellen an, welchen ein Feedback zur Datenqualität benötigt wird, angewendet. Der Auswertungsprozess (2) läuft wie folgt ab:

Stichproben

Zuerst werden mehrere Stichproben der Daten generiert (2, B), auf welchen im späteren Verlauf die Auswertung der Datenqualität erfolgt. Für dieses Prozessmodell wird analog zu [TKS+16] die Strategie „Bag of Little Bootstrap“ (BLB [KTSJ14]) für die Generation der Stichproben (2, B) und die Auswertung der Ergebnisse auf diesen (2, D) verwendet. Bei diesem Verfahren wird zuerst eine große Datenquelle in n-Stichproben ohne Zurücklegen

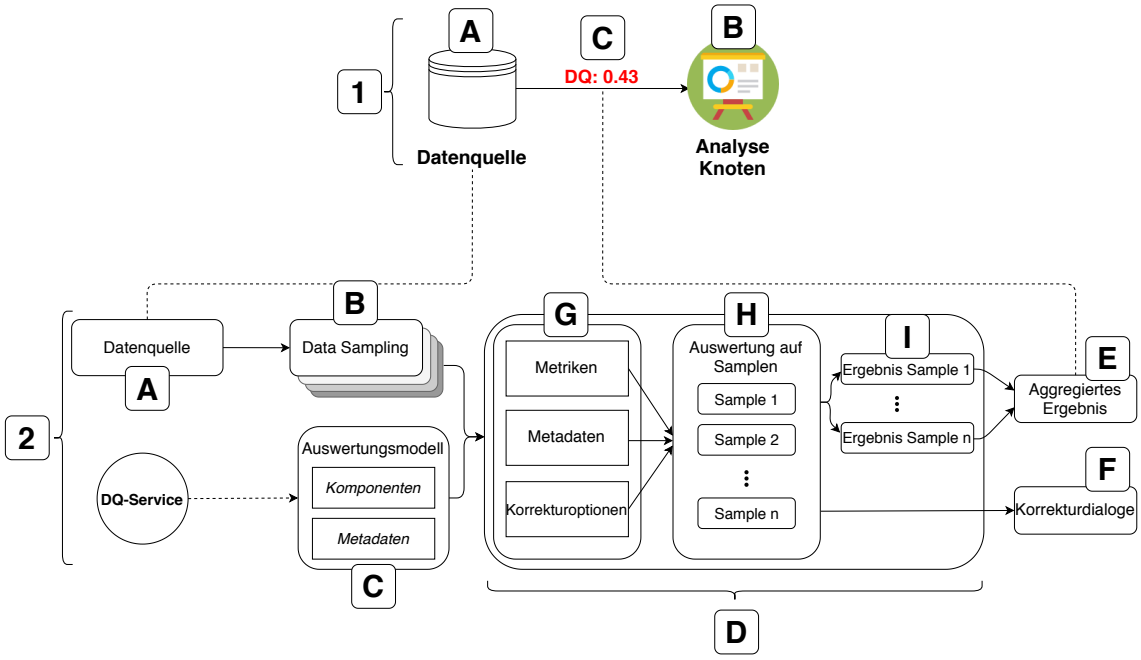

Abbildung 3.13: Prozessmodell (2) zur Auswertung der Datenqualität in der Evaluationsumgebung (adaptiert aus [TKS+16]). Mit exemplarischen Workflow (1) bestehend aus einer Datenquelle (1, A), einem Knoten zur Auswertung der Daten (1, B) sowie einer Auswertungstelle (1, C). Das Prozessmodell (2) selbst setzt sich der Stichproben Erstellung (2, B) und dem Auswertungsmodell (2, C) zusammen, welche für die Auswertung (2, D) verwendet werden. Die für die Auswertung relevanten Komponenten (2, G) werden auf die Stichproben angewendet (2, H) und die Ergebnisse (2, I) zu einem Endergebnis (2, E) aggregiert.

(SoZ) unterteilt [KTSJ14]. Für jede dieser Stichproben (SoZ) werden n-Stichproben mit Zurücklegen (SmZ) erstellt [KTSJ14]. Die Datenqualität wird dann auf diesen Stichproben berechnet und zu einem einzigen Ergebnis (2, E) aggregiert [TKS+16; KTSJ14].

Auswertungsmodell

Für die Auswertung der Stichproben (2, B) wird zusätzlich das von dem Datenqualitäts-Service bereitgestellte Auswertungsmodell (2, C) benötigt. Die im Auswertungsmodell (2, C) enthaltenen Komponenten und Metadaten stellen die Grundlage für die Auswertung der Stichproben dar. Sie beinhalten die zur Auswertung der Datenqualität verwendeten Metriken mit den dazu benötigten Metadaten und der Korrekturoptionen für gefundene Fehler.

Auswertungsprozess

Im Auswertungsprozess (2, D) werden die Metriken unter Verwendung der Metadaten zur Auswertung der Datenqualität der einzelnen Stichproben (2, H) angewendet. Dies resultiert in einem Auswertungsergebnis (jeweils auf verschiedenen Granularitätsebenen) für die Datenqualität der einzelnen Stichproben (2, I). Diese Auswertungsergebnisse beinhalten die Ergebnisse der einzelnen Komponenten des Auswertungsmodells. Diese werden jeweils auf ein durchschnittliches Ergebnis (2, E) je Komponente aggregiert.

Die bei der Auswertung auf den Stichproben gefundenen Probleme werden zur Generierung von Korrekturdialogen (2, F) für den Domänenexperten verwendet. Hierzu werden die in der Komponente, durch welche das Problem identifiziert wurde, enthaltenen Korrekturoptionen verwendet.

Integration der Ergebnisse

Abschließend werden die Auswertungsergebnisse (2, E) und Korrekturdialoge (2, F) in die Benutzeroberfläche integriert (exemplarisch für die Workflow-Ebene in 1, C).

Bei Änderungen im Workflow (beispielsweise durch das Hinzufügen, Ändern, Löschen einer Operation) wird die Datenqualität nur für Operationen, welche von den aus der geänderten Operation resultierenden Daten direkt oder indirekt abhängen, neu berechnet. Die Datenqualität für Operationen im Workflow, welche nicht von den Daten der Operation abhängen, bleibt unverändert und muss daher nicht neu berechnet werden.

4 Implementierung

Das im vorherigen Kapitel 3 vorgestellte Konzept wurde prototypisch in FlexMash (Abschnitt 2.7), einem Datenanalysewerkzeug der Universität Stuttgart, implementiert. Für die Integration der Datenqualitätsüberwachung in FlexMash muss die vorhandene Architektur erweitert werden, sodass die Überwachung der Datenqualität gemäß Kapitel 3 möglich ist.

4.1 Aufbau und Erweiterung

Die Architektur von FlexMash basiert auf einem Client-Server-Modell. Der Domänenexperte interagiert mit einem auf Typescript basierenden Frontend zur Modellierung eines Datenflusses. Für die Modellierung des Datenflusses stehen dem Domänenexperten verschiedene UI-Elemente zur Verfügung. Diese UI-Elemente sowie Informationen über den modellierten Datenfluss (beispielsweise den Status der Ausführung) werden aus dem Backend abgefragt und in das Frontend integriert. Das Backend basiert auf Python und Java und verwaltet die Ausführung des modellierten Datenflusses sowie die verfügbaren Operationen und Datenquellen. Für die Kommunikation zwischen Backend und Frontend wird eine Kombination aus Websockets und dem „Message Queuing Telemetry Transport“-Protokoll (MQTT) verwendet. Die einzelnen Komponenten, aus welchen FlexMash zusammengesetzt ist, werden in Abbildung 4.1 (grau) dargestellt. Ihre Funktion lässt sich wie folgt beschreiben:

Das **Mashup Interface** enthält Menüs und die grundlegende Logik zur Modellierung des Analyseprozesses. Der Funktionsumfang der Mashup-Oberfläche wird UI-Elemente aus dem Backend erweitert.

Der **Partial Executor** verwaltet die Kommunikation der Benutzeroberfläche mit dem Backend und koordiniert die Ausführung des modellierten Analyseprozesses. Für die Koordination der Ausführung erhält der Partial Executor Informationen durch den Dependency Checker und den Data Refresher. Der **Dependency Checker** überprüft, ob die Vorbedingungen zur Ausführung der einzelnen Operationen des modellierten Analyseprozess erfüllt sind, während der **Data Refresher** die Aktualität der Daten überwacht.

Der **Backend Coordinator** dient als zentrale Schnittstelle zwischen allen Komponenten. Er verwaltet den Informationsfluss innerhalb des Backends sowie die Kommunikation zwischen Backend und Frontend.

Im **Intermediate Storage** werden Daten zwischengespeichert. Die **Execution Engine** wendet die im Analysprozess modellierten Operationen auf die Daten an und unterstützt die Ausführung der Operationen auf mehreren Berechnungs-Framworks wie beispielsweise Apache Spark. Das **Service Repository** beinhaltet die für die Modellierung des Analyseprozesses verfügbaren Operationen und speichert die Konfigurationen des Domänenexperten.

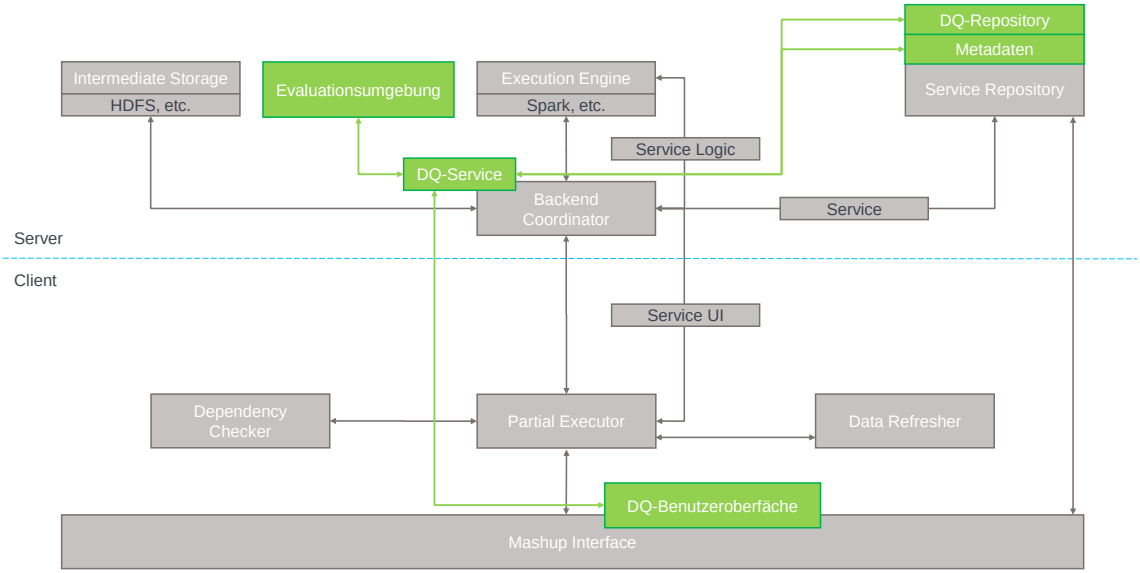

Abbildung 4.1: FlexMash Architektur mit bereits vorhandenen Komponenten (grau) und den Komponenten der Datenqualitätsüberwachung (grün)

4.1.1 Erweiterung mit einer Datenqualitätsüberwachung

Für die Implementierung einer Datenqualitätsüberwachung in FlexMash wurde die bestehende Architektur durch die Datenqualitäts-Komponenten aus Kapitel Kapitel 3 erweitert. Die daraus resultierende Architektur ist in Abbildung 4.1 dargestellt. Die Funktionalität und Notwendigkeit der hinzugefügten Komponenten wird im Folgenden erklärt:

Datenqualitäts-Service: Der DQ-Service (Abschnitt 3.5) wird benötigt, um den korrekten Ablauf der Datenqualitätsüberwachung (Abschnitt 3.1) sicherzustellen. Er kommuniziert daher mit den anderen Datenqualitäts-Komponenten. Für die Koordination der Datenqualitätsüberwachung werden durch den DQ-Service zusätzlich Informationen über den Modellierungsprozess des Domänenexperten (beispielsweise das Hinzufügen einer neuen Datenquelle) und den Status der Execution Engine verwendet. Der DQ-Service wird daher bei der Implementierung in den Backend Coordinator integriert, da dieser aufgrund seiner zentralen Stellung in der FlexMash Architektur Zugriff auf die vom DQ-Service verwendeten Informationen des Analyseprozesses hat.

Datenqualitäts-Benutzeroberfläche: Damit der Domänenexperte über die Datenqualität informiert wird und die Interaktionen mit der Datenqualität gewährleistet werden kann, muss die Benutzeroberfläche (Abschnitt 3.4) die entsprechenden Dialoge (Abschnitt 3.4) beinhalten. Die DQ-Benutzeroberfläche erweitert daher die vorhandene Benutzeroberfläche von FlexMash durch die benötigten Bestandteile der Datenqualität. Die angezeigten Datenqualitäts-Informationen erhält die Benutzeroberfläche durch den DQ-Service.

Evaluationsumgebung: Die Evaluationsumgebung (Abschnitt 3.6) wird zur Auswertung der Datenqualität benötigt. Für die Auswertung der Datenqualität benötigt die Evaluationsumgebung Informationen (beispielsweise Auswertungslogik, Metadaten und Korrekturoptionen). Die Evaluationsumgebung kommuniziert daher mit dem DQ-Service. Die Ergebnisse der Auswertung werden vom DQ-Service an die DQ-Benutzeroberfläche weitergeleitet.

Datenqualitäts-Repository: Im DQ-Repository (Abschnitt 3.3) wird die Programmlogik zur Auswertung der Datenqualitäts-Dimensionen mit den jeweils zur Auswertung benötigten Parametern, sowie den entsprechenden Korrekturoptionen gespeichert. Der DQ-Service kommuniziert mit dem DQ-Repository zur Abfrage der auswertbaren Dimensionen und nützlichen Parameter.

Metadaten-Repository: Die Metadaten Komponente (Abschnitt 3.2) wird zur Speicherung der Metadaten der verwendeten Datenquellen benötigt. Die in der Komponente gespeicherte Metadaten können vom DQ-Service abgefragt oder bei Eingabe durch den Domänenexperte hinzugefügt und geändert werden. Für FlexMash genügt die Implementierung der Metadaten Komponente, da FlexMash bereits eigenständig eine Funktionalität zur Verwaltung vorhandener Datenquellen bietet. Es muss daher nicht das vollständige Datenquellen-Repository implementiert werden.

4.2 Prototypische Implementierung

Auf Basis der obigen Architektur wurde prototypisch eine Datenqualitätsüberwachung in FlexMash integriert. Die Integration orientiert sich an den in Kapitel 3 definierten Richtlinien und beachtet daher auch die Anforderungen des Domänenexperten (Abschnitt 1.1) an die Integration einer Datenqualitätsüberwachung. Von FlexMash wird eine englischsprachige Benutzeroberfläche eingesetzt. Die Erweiterung der Benutzeroberfläche wird daher ebenfalls durch die Verwendung englischsprachiger Elemente vollzogen.

4.2.1 Integration in den Analyseprozess

Für die Integration der Datenqualität in den Analyseprozess des Domänenexperten wurde die bereits in FlexMash vorhandene Benutzeroberfläche zur Modellierung von Datenflüssen erweitert. Die Integration der Datenqualität in die Benutzeroberfläche wurde gemäß Abschnitt 3.4 auf mehreren Granularitätsebenen vollzogen.

Workflowebene

Bei der Integration der Datenqualität auf Workflowebene (Abschnitt 3.4) wurden die Kanten, welche FlexMash für die Modellierung von Datenflüssen zwischen mehreren Komponenten verwendet um einen Datenqualitäts-Indikator erweitert (Abbildung 4.2). Der Domänenexperte wird bei der Verwendung jeder Komponente (beispielsweise zur Transformation der Daten oder zur Auswertung der Daten) über die Qualität der Daten, von welcher die Ausführung der Komponente abhängt, informiert. Er kann so den Einfluss seiner Operationen auf die Datenqualität nachvollziehen.

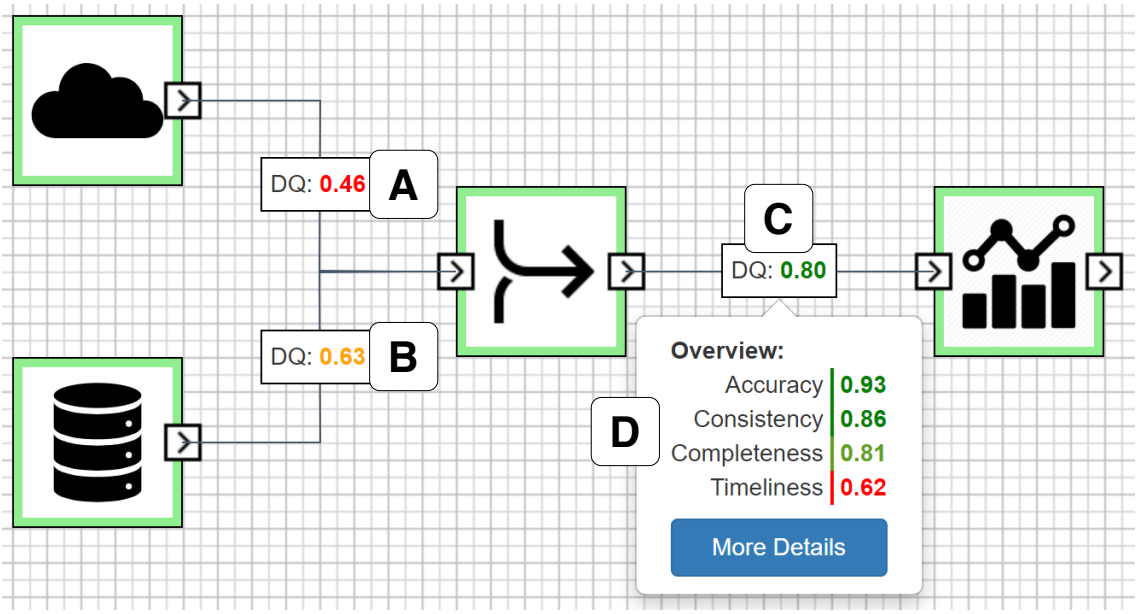

Abbildung 4.2: In FlexMash modellierter Datenfluss mit integrierten Datenqualitäts-Indikatoren (A, B, C), sowie einem ausführlicheren Feedback auf Dimensionsebene (D)

Ein Beispiel stellt der in FlexMash modellierte Datenfluss aus Abbildung 4.2 in welchem der Domänenexperte Cloud-Daten mit schlechter Qualität (A) und Daten aus einer Datenbank mit mäßiger Qualität (B) zusammenführt. Der Domänenexperte wird in (C) über die (positive) Auswirkung seiner Operation informiert und kann anschließend eine informierte Entscheidung treffen, ob die resultierende Datenqualität für die Verwendung der Daten in einer darauffolgenden Analyse-Komponente ausreichend ist.

Dimensionsebene

Damit der Domänenxperte die Datenqualität von den im Datenfluss Diagramm enthaltenen Komponenten verwendeten Daten präziser bewerten kann wurde (nach Abschnitt 3.4) ebenfalls ein Feedback über die Datenqualität auf **Dimensionsebene** (Abbildung 4.2, D) in den Datenfluss implementiert. Der Domänenexperte kann hierfür mit den angezeigten Datenqualitätsindikatoren (A, B, C) interagieren um einen Dialog, der die Bewertungen der einzelnen Datenqualitäts-Dimensionen enthält (D) zu betrachten. Dadurch wird eine genauere Identifikation von Datenqualitätsproblemen ermöglicht.

Der Domänenexperte kann beispielsweise in Abbildung 4.2 nach dem Zusammenfügen der Daten die Zusammensetzung des Ergebnisses betrachten (D) und die Aktualität der Daten als potenzielles Problem identifizieren.

Für eine möglichst einfache Identifikation von Problemen werden die Auswertungsergebnisse der beiden Granularitätsebenen farbcodiert. Problemstellen werden dem Domänenexperten in rot angezeigt und akzeptable bis gute Ergebnisse in Orange und Grün. Der Farbverlauf kann an die Bedürfnisse des Domänenexperten angepasst werden.

Attributsebene

Dem Domänenexperten wird im Dialog auf der Datenqualität auf Dimensionsebene (Abbildung 4.2) zusätzlich die Option geboten einen detaillierten Bericht über die Datenqualität auf **Attributsebene** (Abschnitt 3.7) aufzurufen. Für den detaillierten Bericht (Abbildung 4.3) wird aufgrund der im Vergleich zu dem vorherigen Dialog deutlich erhöhten Informationsmenge ein eigenständiges Fenster geöffnet. Hierfür wird das in FlexMash integrierte Fensterverwaltungssystem verwendet, um einen neuen „Tab“ anzulegen (siehe Abbildung 4.3, C). Der Domänenexperte kann so flüssig zwischen der Datenflussmodellierung und der Datenqualitäts-Übersicht wechseln.

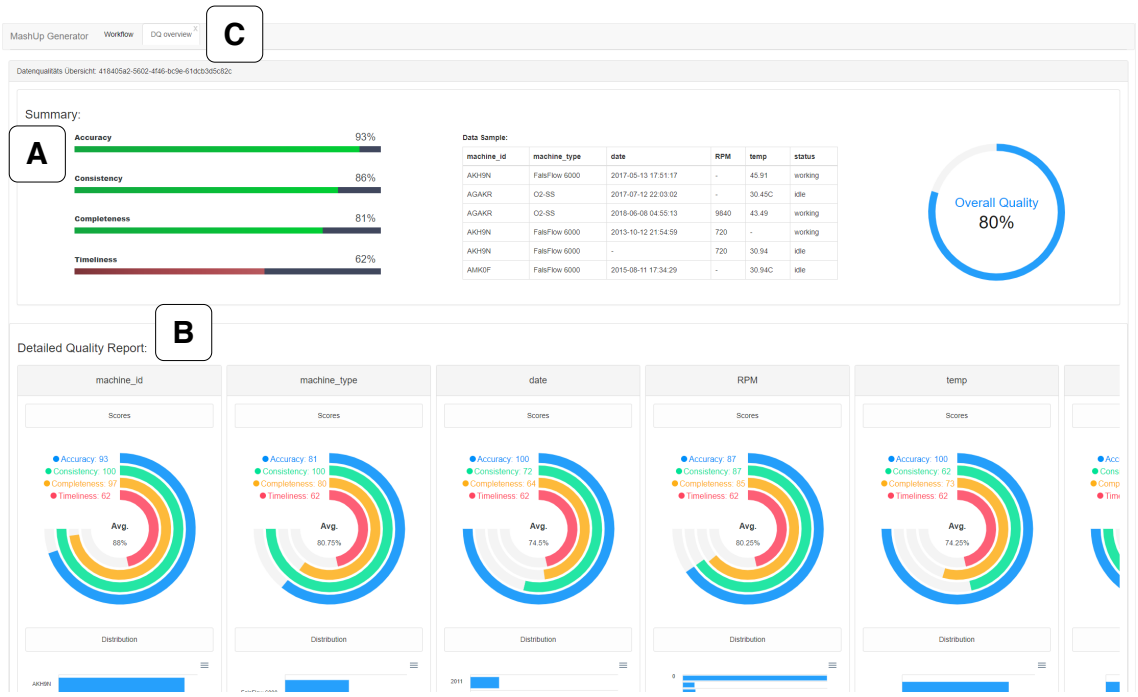

Abbildung 4.3: Ausschnitt aus dem Datenqualitäts-Übersichtsfenster auf Attributsebene mit einer Zusammenfassung der Datenqualität (A) und einer ausführlichen Auswertung der einzelnen Dimensionen (B), sowie die Verwendung verschiedener Tabs (C)

Das Übersichtsfenster (Abbildung 4.3) setzt sich aus einer Zusammenfassung der Datenqualität (A), sowie einem detaillierten Bericht über die Datenqualität der einzelnen Attribute (B) zusammen. Für eine vereinfachte und anschaulichere Darstellung der Ergebnisse (Abschnitt 3.4) wurden gezielt Visualisierungen verwendet. Zur Visualisierung der Ergebnisse wurde die Javascript Bibliothek „APEXCHARTS“[1] verwendet. Diese ermöglicht unter anderem die Aktualisierung bestehender Visualisierungen, wodurch Änderungen in der Datenqualität einfacher in den bestehenden Visualisierungen reflektiert werden können.

[1] https://apexcharts.com/

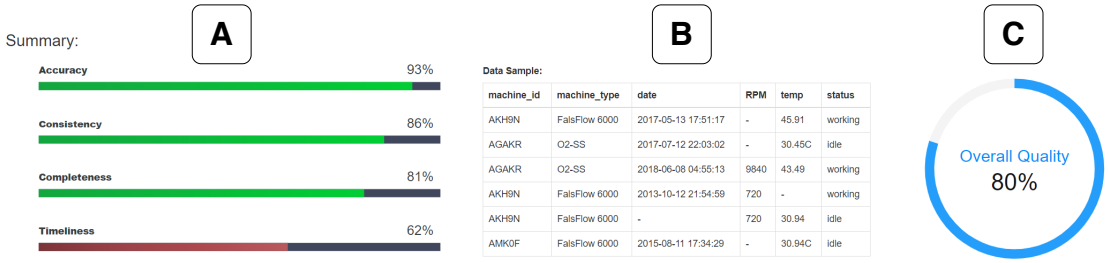

Abbildung 4.4: Zusammenfassung aus der Datenqualitätsübersicht mit den einzelnen Dimensionen (A), einem Ausschnitt der Daten (B), sowie der insgesamten Datenqualität (C)

Abbildung 4.5: Detaillierter Report über die Datenqualität auf Attributsebene mit den jeweiligen Elementen der einzelne Attribute (A), sowie den einzelnen Bestandteilen dieser (B, C, E)

Die im Übersichtsfenster enthaltene Zusammenfassung der Datenqualität beinhaltet eine Visualisierung der Datenqualitäts-Dimensionen in Form von „Progress-Bars“ (Abbildung 4.4, A) sowie einen Ausschnitt der Daten (B), welcher dem Domänenexperten einen Kontext für die attributsbezogenen Informationen stellt und die durchschnittliche Datenqualität der Dimensionen (C). Die Zusammenfassung vereinfacht das Verständnis des detaillieren Berichtes (Abbildung 4.3, B) und verhindert den unnötigen Wechsel auf die Datenfluss-Ansicht für die Betrachtung aggregierter Werte.

Der detaillierte Datenqualitäts-Bericht (Abbildung 4.5) informiert den Domänenexperten über die Datenqualität der einzelnen Attribute (Abschnitt 3.4). Für die Erstellung des Reports wird für jedes Attribut der Daten eine separates grafisches Element erstellt (A). Diese Elemente setzen sich den Auswertungsergebnissen der Datenqualitätsdimensionen für die jeweiligen Attribute (B) einem Diagramm zur Darstellung der Werteverteilung (C), einer Reihe statistischer Informationen über

die Daten (D) sowie den automatisierten Korrekturoptionen für die gefundenen Datenqualitätsprobleme zusammen. Für die Erstellung der einzelnen Bestandteile werden die Informationen über die Datenqualität aus dem Backend verwendet, um die einzelnen Bestandteile der Elemente (A) aufzubauen. Im Folgenden wird die Zusammensetzung der Elemente näher betrachtet:

Die Abbildung 4.6 stellt die detaillierte Datenqualitäts-Ansicht der beiden Attribute „date“ und „RPM“ dar. Dem Domänenexperten wird hierfür eine kompakte Visualisierung der Auswertungsergebnisse der verschiedenen Dimensionen angezeigt. Zur Visualisierung der Dimensionen wird ein „Radial Bar Chart“ (A) verwendet. Dem Domänenexperten wird so eine Übersicht über die Probleme der verschiedenen Attribute gestellt. Für ein besseres Verständnis der Daten wird ihm die Werteverteilung des Attributes (B), sowie eine Bandbreite statistischer Maße (C) gestellt. Der Domänenexperte kann diese verwenden, um die Auswertung der einzelnen Dimensionen besser zu verstehen. Er kann beispielsweise die verbesserungswürdige Vollständigkeit der Daten des „date“ Attributs mit der hohen Anzahl an *Null* Werten innerhalb des Attributes assoziieren. Eine weitere Erklärung für die Probleme der einzelnen Attribute, sowie die Möglichkeit zur Korrektur dieser wird dem Domänenexperten durch die Implementierung der Korrekturoptionen (Abschnitt 3.4) geboten (D). Bei der Implementierung der Korrekturoptionen wurden aufgrund der technischen Komplexität einiger Funktionen (wie der Vorschau der Veränderung der Werte innerhalb der Daten, sowie der vorhersage der Auswirkung auf die Datenqualität) nicht alle Funktionen aus Abschnitt 3.4 implementiert. Die vorhandene Implementierung (D) beinhaltet eine Beschreibung des Problems, sowie eine Aufzählung der vom System automatisiert ausführbaren Korrekturmaßnahmen mit den eventuell benötigten Dialogen zur Anwendung. Die derzeitige Implementierung beinhaltet Korrekturoptionen für eine Reihe von Problemen (fehlende Werte, verletzte Regeln, unterschiedliche Datentypen, veraltete Werte und verschiedene Formate). Der Domänenexperten wird anhand der Korrekturoptionen des Attributes „RPM“ (Abbildung 4.6, E) beispielsweise auf Werte hingewiesen welche eine von ihm definierte Regel verletzen. Er kann die Werte, welche seine Regel verletzen automatisiert entfernen oder seine Regel mithilfe eines Dialoges anpassen.

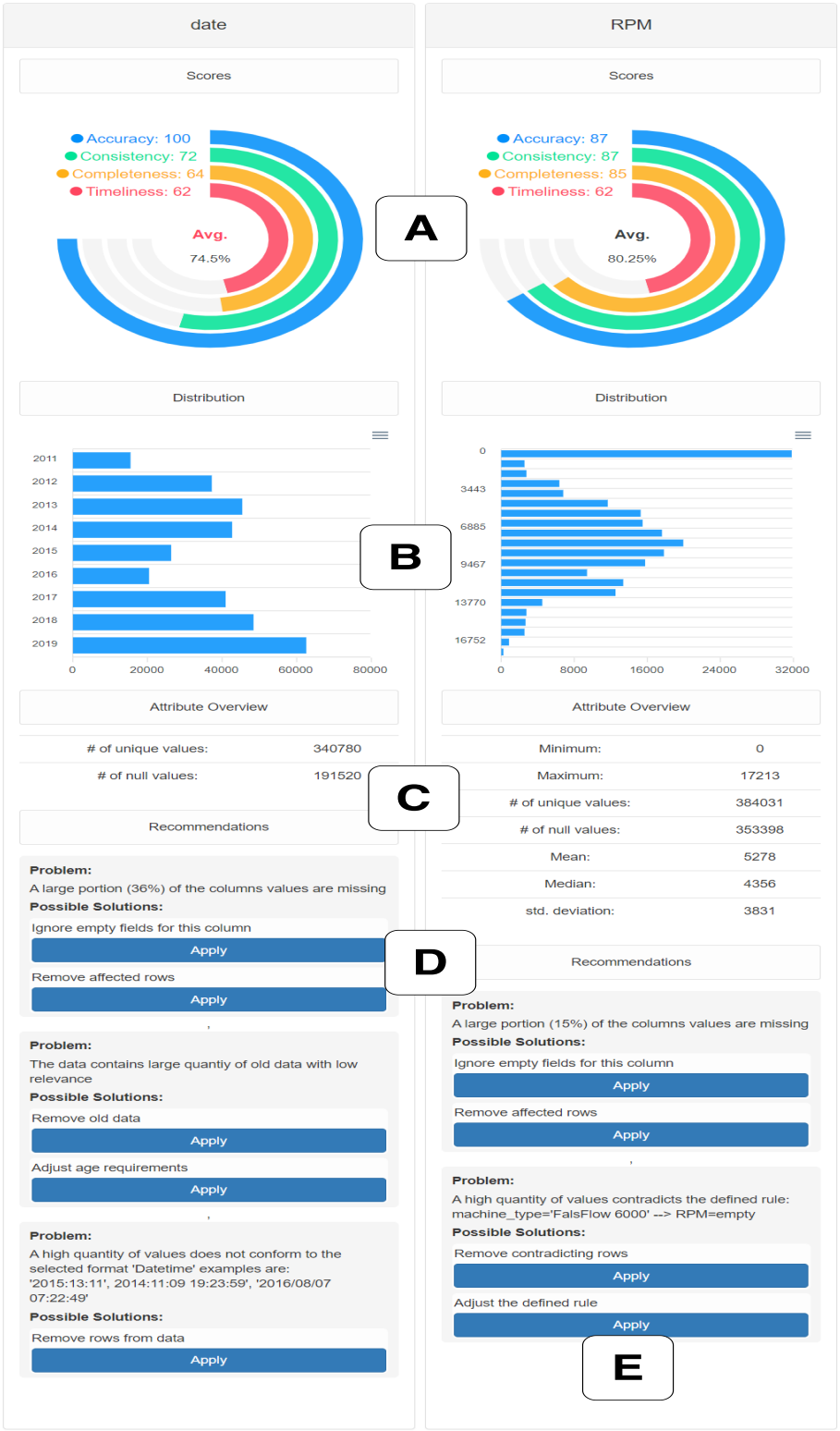

Abbildung 4.6: Detaillierter Datenqualitäts-Bericht der beiden Attribute ’date’ und ’RPM’ mit den visualisierten Ergebnissen der einzelnen Dimensionen (A), der Werteverteilung der Attribute (B) die statistische Beschreibung der Attribute (C), sowie Korrekturoptionen zur Verbesserung von Problemen (D, E)

Metadaten

Für den korrekten Ablauf der Datenqualitätsüberwachung werden zusätzlich Metadaten benötigt. Für den Domänenexperten wurde daher ein Dialog zur einfachen Bereitstellung (Abschnitt 3.4) von Metadaten implementiert. Um die Komplexität der Implementierung zu reduzieren, beschränkt sich die Implementierung auf ausgewählte Metadaten. Der resultierende Dialog ist in Abbildung 4.7 dargestellt. Dieser bietet dem Domänenexperten eine Übersicht über die Daten, für welche er Metadaten angibt (A). Er kann das Attribut, für welches er Metadaten ergänzen möchte auswählen (B) und dieses wird in den Daten farbig hervorgehoben (C). Dem Domänenexperten werden nach der Auswahl des Attributes Dialoge zur Ergänzung von Metadaten angezeigt (D, E, F, G, H). Derzeitig implementiert ist ein Dialog zur Auswahl des Datentypen (D), die Möglichkeit doppelt vorkommende Werte zu erlauben/verbieten (F), die Option einen Wertebereich für die Werte des Attributes anzugeben (G) und die Option benutzerdefinierte Regeln auf den Daten zu definieren (H). Für die Angabe der Geschäftsregeln (H) wurde aus Implementierungsgründen nicht die DM Notation verwendet.

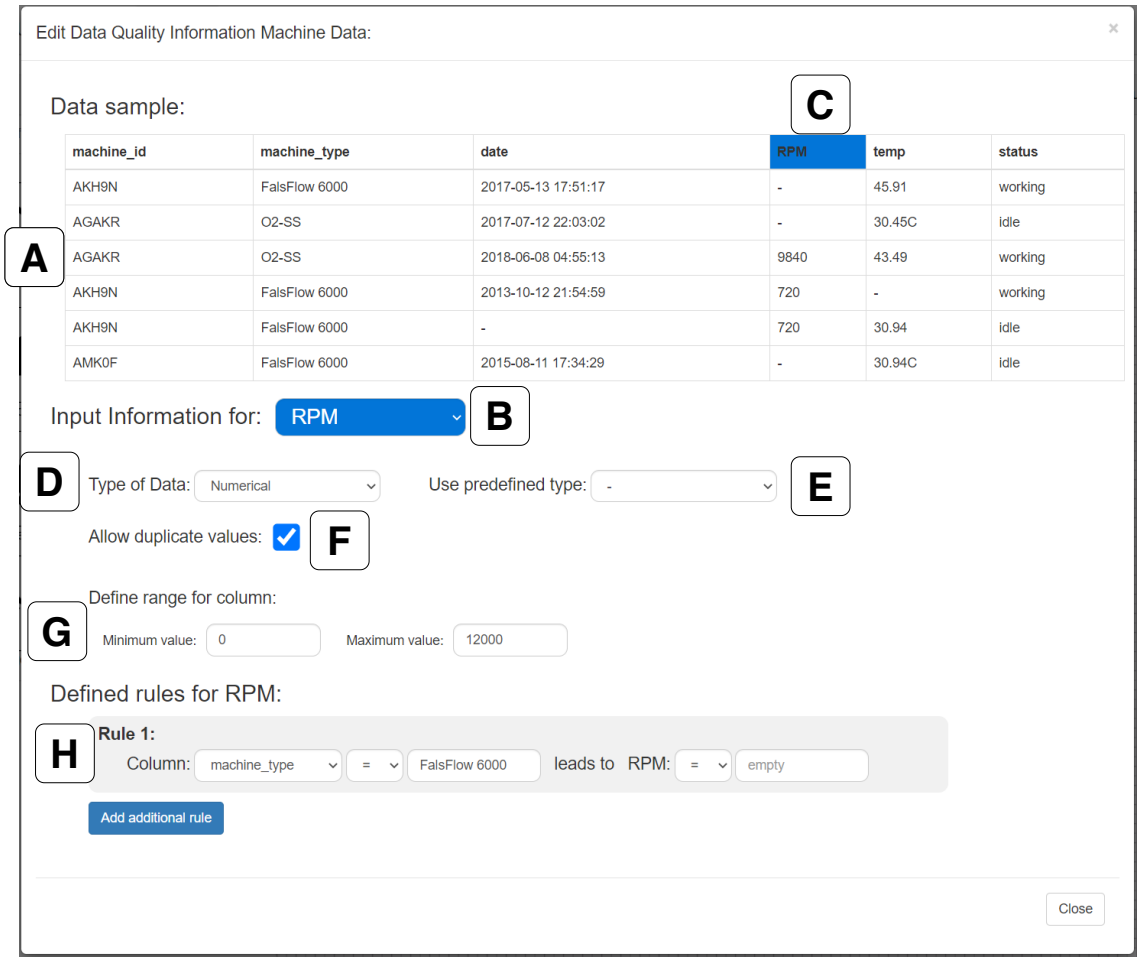

Abbildung 4.7: Dialog zur Eingabe von Metadaten mit einer Übersicht über die Daten (A), dem ausgewählten Attribut (B, C) und Dialogen zur Angabe von Metadaten (D-H)

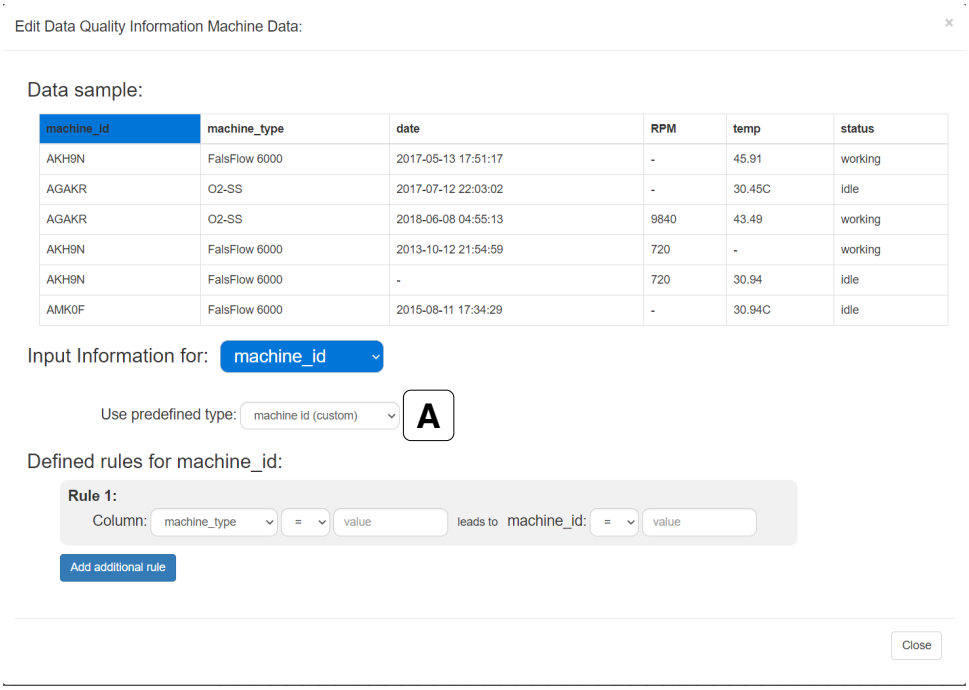

Abbildung 4.8: Verwendung einer Vorlage (A) für ein Attribut

Der Domänenexperte hat in Abbildung 4.7 exemplarisch die Metadaten für das Attribut „RPM“ eingetragen. Das Attribut enthält nicht eindeutige Zahlenwerte. Der Domänenexperte wählt in (D) daher den entsprechenden Datentyp aus und erlaubt doppelte Werte (F). Der Domänenexperte weiß, dass die Maschinen seiner Firma nicht mehr als 12000 RPM leisten können. Er definiert daher einen Wertebereich für das RPM Attribut mit einem Maximum von 12000 (G). Aus seiner Expertise weiß der Domänenexperte außerdem, dass die Maschine „FalsFlow 6000“ eine Laser-Maschine ist und daher keine RPM besitzt. Er definiert daher die Regel machine_type=„FalsFlow 600“ —> RPM=null.

Der implementierte Dialog zur Angabe von Metadaten beinhaltet außerdem einen Dialog zur Auswahl vordefinierter Metadaten (E). Ein IT-Experte kann seine Expertise nutzen, um Vorlagen für häufig verwendete Attribute vorab zu erstellen. Ein Beispiel für die Verwendung dieser Vorlagen stellt Abbildung 4.8 dar. Der Domänenexperte hat hier für das Attribute „machine_id“ die entsprechende Vorlage ausgewählt (A). Da vom System keine weiteren Metadaten benötigt werden, werden die restlichen Dialoge ausgeblendet.

In der prototypischen Implementierung ist keine Benutzeroberfläche für den IT-Experten enthalten. Er kann derzeit Metadaten und Typen direkt in der Datenbank definieren.

Bewertung der Integration in den Analyseprozess

Durch die Integration der Datenqualitätsüberwachung in FlexMash auf den Granularitätsebenen „Workflowebene“, „Dimensionsebene„ und „Atrributsebene“ wird dem Domänenexperten ein umfassendes Feedback über die Datenqualität auf mehreren Ebenen ermöglicht. Die Implementierung der Datenqualitätsüberwachung auf Datenebene (Abschnitt 3.4), um den Domänenexperten bei

direkten Arbeiten auf den Daten anhand der Qualität einzelner Tupel und Werte zu unterstützen, wurde aufgrund der technischen Komplexität vernachlässigt. Aufgrund der direkten Integration der Datenqualitätsüberwachung in die Modellierung des Datenflusses wird der Domänenexperte während des vollständigen Analyseprozesses durch die Datenqualitätsüberwachung unterstützt. Für eine verständliche Darstellung der Auswertungsergebnisse wurde eine Reihe von Maßnahmen (etwa der Visualisierung und der Erklärung der Korrekturoptionen) implementiert. Automatisierte Korrekturoptionen für die gefundenen Probleme sind in grundlegender Form implementiert. Die Implementierung des Feedbacks über die Datenqualität erfüllt daher die Kriterien für die Implementierung der Benutzeroberfläche aus Abschnitt 3.4 (K1-K5).

5 Zusammenfassung und Ausblick

Aufgrund der exponentiell steigenden Datenmenge werden neue Lösungen für die Analyse der Daten benötigt. Automatisierte Ansätze zur Analyse der Daten sind als Lösung unzureichend, da das domänenspezifische Wissen des Anwenders nicht automatisiert werden kann. Dieses ist jedoch für das Ergebnis der Analyse entscheidend. Es wird daher eine eigenständige interaktive Analyse der Daten durch den Anwender benötigt. Bei der eigenständigen interaktiven Analyse von Daten durch Domänenexperten stellen fehlerhafte Daten ein Problem dar. Daraus ergibt sich die Herausforderung den Domänenexperten während der interaktiven Analyse mithilfe der Datenqualität zu unterstützen, um den Einfluss fehlerhafter Daten auf das Ergebnis der Analyse zu minimieren.

In dieser Arbeit wurden die Anforderungen des Domänenexperten an eine effektive Integration der Datenqualität in den Analyseprozess identifiziert. Anhand dieser Anforderungen konnte gezeigt werden, dass vorhandene Werkzeuge dem Domänenexperten keine adäquate Unterstützung während des Analyseprozesses bieten.

Es wurde daher ein Konzept zur Unterstützung des Domänenexperten während des Analyseprozesses ausgearbeitet. Diesbezüglich wurde ein zu den Anforderungen des Domänenexperten konformes Prozessmodell zur Integration der Datenqualitätsüberwachung in den interaktiven Analyseprozess eines Domänenexperten eingeführt. Die Funktionsweise des Prozessmodells wurde anhand eines Beispiels dargestellt. Damit die effektive Unterstützung des Domänenexperten sichergestellt wird, wurden die Eigenschaften und Funktionsweisen der im Prozessmodell verwendeten Komponenten aufgezeigt und Richtlinien für die Implementierung der Komponenten definiert. In diesem Zug wurde ein modular erweiterbares dimensionsbasiertes Modell zur Evaluation der Datenqualität mithilfe von Metriken vorgestellt. Die Evaluation der vier wichtigsten Datenqualitäts-Dimensionen durch dieses Modell wurde demonstriert und die Auswahl der verwendeten Metriken begründet. Für eine möglichst effektive Evaluation der Datenqualität wurde auch die Rollenaufteilung zwischen Domänenexperten und IT-Experten dargelegt. Dabei lag der Fokus vor allem auf der Verwendung der unterschiedlichen Expertisen der beiden Nutzergruppen zur Verbesserung des Evaluations-Ergebnisses. Die Integration der aus der Evaluation der Datenqualität resultierenden Informationen über die Datenqualität in den Analyseprozess des Domänenexperten wurde auf mehreren Granularitätsebenen spezifiziert und anhand konkreter Beispiele verdeutlicht. Des Weiteren wurden Interaktionsmöglichkeiten des Domänenexperten mit den Evaluations-Ergebnissen identifiziert. Abschließend wurde eine Methodik für die Anwendung des Evaluation-Modells auf großen Datenmengen durch eine Stichproben-basierte Auswertung der Metriken eingeführt.

Das in dieser Arbeit entstandene Prozessmodell für die Unterstützung des Domänenexperten durch die Integration einer Datenqualitätsüberwachung in den Analyseprozess erfüllt alle identifizierten Anforderungen des Domänenexperten an die Integration einer Datenqualitätsüberwachung. Es kann daher als Grundlage für die Integration einer Datenqualitätsüberwachung in den interaktiven Analyseprozess des Domänenexperten verwendet werden.

Das entwickelte Konzept wurde prototypisch in das Mashup-Werkzeug „FlexMash“ zur Modellierung von Datenflüssen der Universität Stuttgart implementiert. Bei der Implementierung wurde die bereits vorhandene Architektur von FlexMash durch die einzelnen Komponenten des erarbeiteten Prozessmodells erweitert.

Ausblick

Das Konzept wurde entwickelt, um den Domänenexperten während des Analyseprozesses durch die Integration einer Datenqualitätsüberwachung zu unterstützen. Es stellt sich daher die Frage, welchen Effekt die Integration der Datenqualitätsüberwachung auf das Analyseergebnis des Domänenexperten hat. Um diese Frage überzeugend beantworten zu können wäre eine Nutzerstudie notwendig, in welche die Ergebnisse mit und ohne die Unterstützung der Datenqualitätsüberwachung gegenüberstellt. Für die Durchführung der Nutzerstudie könnte eine (weiterentwickelte) Form der prototypischen Implementierung verwendet werden.

Zusätzlich kann das Konzept durch die Integration weiterer Funktionalitäten erweitert werden. Hierfür sind eine Vielzahl von Erweiterungsmöglichkeiten möglich. Es ist etwa die Verwendung von Data Provenance denkbar, um dem Domänenexperten eine noch einfachere Identifikation von Fehlerquellen zu ermöglichen. Es ist außerdem denkbar, zusätzlich zur Datenqualität auch die Qualität der verfügbaren Metadaten zu überwachen.

Literaturverzeichnis

[AA19] M. Al-Mekhlal, A. Ali Khwaja. „A Synthesis of Big Data Definition and Characteristics“. In: *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*. IEEE, Aug. 2019. DOI: 10.1109/CSE/EUC.2019.00067. URL: https://ieeexplore.ieee.org/document/8919591/ (zitiert auf S. 17).

[AG09] M. d. P. Angeles, F. García-Ugalde. „A Data Quality Practical Approach“. In: *International Journal on Advances in Software* 2.3 (2009), S. 259–274. URL: http://www.iariajournals.org/software/ (zitiert auf S. 32).

[ASWW18] O. Azeroual, G. Saake, J. rgen Wastl, J. Wastl JuergenWastl. „Data measurement in research information systems: metrics for the evaluation of data quality“. In: *Scientometrics* 115 (2018). URL: https://doi.org/10.1007/s11192-018-2735-5 (zitiert auf S. 18, 19, 21, 24, 48, 51).

[AW14] P. Alpar, S. Winkelsträter. „Assessment of data quality in accounting data with association rules“. In: *Expert Systems with Applications* 41.5 (Apr. 2014), S. 2259–2268. DOI: 10.1016/j.eswa.2013.09.024 (zitiert auf S. 25, 48, 49).

[BHM17] M. Behringer, P. Hirmer, B. Mitschang. „Towards Interactive Data Processing and Analytics Putting the Human in the Center of the Loop“. In: (2017). DOI: 10.5220/0006326300870096 (zitiert auf S. 11).

[BHM18] M. Behringer, P. Hirmer, B. Mitschang. „A Human-Centered Approach for Interactive Data Processing and Analytics“. In: Bd. 321. Springer Verlag, Apr. 2018, S. 498–514. DOI: 10.1007/978-3-319-93375-7_23 (zitiert auf S. 14, 25).

[BKBJ14] B. Behkamal, M. Kahani, E. Bagheri, Z. Jeremic. „A Metrics-Driven Approach for Quality Assessment of Linked Open Data“. In: *J. Theor. Appl. Electron. Commer. Res.* 9.2 (Mai 2014), S. 64–79. DOI: 10.4067/S0718-18762014000200006. URL: https://doi.org/10.4067/S0718-18762014000200006 (zitiert auf S. 23).

[BM09] R. H. Blake, P. Mangiameli. „Evaluating the Semantic and Representational Consistency of Interconnected Structured and Unstructured Data“. In: *AMCIS 2009 Proceedings*. 2009. URL: http://aisel.aisnet.org/amcis2009/126 (zitiert auf S. 18–20, 54).

[BM11] R. Blake, P. Mangiameli. „The Effects and Interactions of Data Quality and Problem Complexity on Classification“. In: *J. Data and Information Quality* 2.2 (Feb. 2011). DOI: 10.1145/1891879.1891881. URL: https://doi.org/10.1145/1891879.1891881 (zitiert auf S. 18, 23).

[BS06] C. Batini, M. Scannapieca. *Data Quality*. Data-Centric Systems and Applications. Springer Berlin Heidelberg, 2006, S. 235. DOI: https://doi.org/10.1007/3-540-33173-5. URL: http://link.springer.com/10.1007/3-540-33173-5 (zitiert auf S. 19, 20, 23, 24, 42, 52).

[BWPT98] D. Ballou, R. Wang, H. Pazer, G. K. Tayi. „Modeling information manufacturing systems to determine information product quality“. In: *Management Science* 44.4 (1998), S. 462–484. DOI: 10.1287/mnsc.44.4.462 (zitiert auf S. 22, 56, 57).

[CDFS11] B. Carlo, B. Daniele, C. Federico, G. Simone. „A Data Quality Methodology for Heterogeneous Data“. In: *International Journal of Database Management Systems* 3.1 (Feb. 2011), S. 60–79. DOI: 10.5121/ijdms.2011.3105. URL: http://www.airccse.org/journal/ijdms/papers/3111ijdms05.pdf (zitiert auf S. 24, 32).

[Cod90] E. Codd. *The Relational Model for Database Management*. Version 2. Addison-Wesley Publishing Company, 1990 (zitiert auf S. 19, 20, 54, 55).

[CR19] C. Cichy, S. Rass. „An overview of data quality frameworks“. In: *IEEE Access* 7 (2019). DOI: 10.1109/ACCESS.2019.2899751 (zitiert auf S. 18, 35).

[CZ15] L. Cai, Y. Zhu. „The Challenges of Data Quality and Data Quality Assessment in the Big Data Era“. In: *Data Science Journal* 14.0 (Mai 2015), S. 2. DOI: 10.5334/dsj-2015-002. URL: http://datascience.codata.org/article/10.5334/dsj-2015-002/ (zitiert auf S. 18, 32, 48).

[DM14] F. Daniel, M. Matera. *Mashups*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. ISBN: 978-3-642-55048-5. DOI: 10.1007/978-3-642-55049-2. URL: http://link.springer.com/10.1007/978-3-642-55049-2 (zitiert auf S. 18).

[Eck09] W. W. Eckerson. *TDWI Checklist report Self-Service BI*. Techn. Ber. tdwi, 2009. URL: https://www.microstrategy.com/Strategy/media/downloads/white-papers/TDWI_Self-Service-BI.pdf (zitiert auf S. 11).

[ES07] A. Even, G. Shankaranarayanan. „Utility-driven assessment of data quality“. In: *ACM SIGMIS Database* 38 (Mai 2007). DOI: 10.1145/1240616.1240623. URL: http://portal.acm.org/citation.cfm?doid=1240616.1240623 (zitiert auf S. 21).

[Eve05] S. G. Even Adir. „Value-Driven Data Quality Assessment“. In: *Proceedings of the 2005 International Conference on Information Quality {(MIT} {ICIQ} Conference)*. 2005 (zitiert auf S. 23, 24, 56, 57).

[Gar97] E. Gardyn. „A Data Quality Handbook for a Data Warehouse“. In: *IQ*. 1997 (zitiert auf S. 48).

[HBSA18] Heinrich, Bernd, Schiller, Alexander. „Assessing Data Quality-A Probability-based Metric for Semantic Consistency“. In: *Decision Support Systems* 110 (2018), S. 95–106. DOI: 10.1016/j.dss.2018.03.011. URL: https://www.sciencedirect.com/science/article/pii/S0167923618300599 (zitiert auf S. 20, 54).

[HD00] J. A. Hoxmeier, C. DiCesare. „System Response Time and User Satisfaction: An Experimental Study of Browser-based Applications“. In: *AMCIS 2000 Proceedings*. 2000. URL: https://aisel.aisnet.org/amcis2000/347 (zitiert auf S. 70).

[HDV17] F. Hasić, J. De Smedt, J. Vanthienen. „Towards Assessing the Theoretical Complexity of the Decision Model and Notation (DMN)“. In: *International Working Conference on Business Process Modeling, Development and Support (BPMDS)*. Juni 2017 (zitiert auf S. 49).

[HHK+18] B. Heinrich, D. Hristova, M. Klier, A. Schiller, M. Szubartowicz. „Requirements for Data Quality Metrics“. In: *Journal of Data and Information Quality* 9 (Jan. 2018). DOI: 10.1145/3148238. URL: http://dl.acm.org/citation.cfm?doid=3155015.3148238 (zitiert auf S. 21, 45–47).

[Hin02] H. Hinrichs. „Datenqualitätsmanagement in Data Warehouse-Systemen“. Diss. Universität Oldenburg, 2002 (zitiert auf S. 18, 19, 21, 23, 24, 47).

[Hir15] P. Hirmer. „Flexmash – flexible data mashups based on pattern-based model transformation“. In: *Communications in Computer and Information Science*. Bd. 591. Springer Verlag, 2015, S. 12–30. DOI: 10.1007/978-3-319-28727-0_2 (zitiert auf S. 11).

[HK11] B. Heinrich, M. Klier. „Assessing data currency - A probabilistic approach“. In: *Journal of Information Science* 37 (2011). DOI: 10.1177/0165551510392653 (zitiert auf S. 20, 22, 56, 57).

[HKG12] B. Heinrich, M. Klier, Q. Görz. „Ein metrikbasierter Ansatz zur Messung der Aktualität von Daten in Informationssystemen“. In: *Zeitschrift für Betriebswirtschaft* 82 (Nov. 2012). DOI: 10.1007/s11573-012-0623-7. URL: http://link.springer.com/10.1007/s11573-012-0623-7 (zitiert auf S. 18, 20).

[HKK08] B. Heinrich, M. Kaiser, M. Klier. „Does the EU Insurance Mediation Directive Help to Improve Data Quality? A Metric- Based Analysis“. In: *ECIS*. 2008 (zitiert auf S. 19, 23).

[HKK09] B. Heinrich, M. Kaiser, M. Klier. „A Procedure to Develop Metrics for Currency and its Application in CRM“. In: *ACM J. Data Inform. Quality 1, 1, Article* 5 (2009). DOI: 10.1145/1515693.1515697. URL: http://doi.acm.org/10.1145/1515693.1515697 (zitiert auf S. 21).

[HKKW07] B. Heinrich, M. Kaiser, M. Klier, J. Webster. „How to Measure Data Quality? - A Metric-Based Approach“. In: *Proceedings of the 28th International Conference on Information Systems*. 2007. URL: https://epub.uni-regensburg.de/23633/ (zitiert auf S. 21).

[HMHN07] J. Hipp, M. Müller, J. Hohendorff, F. Naumann. „Rule-Based Measurement Of Data Quality In Nominal Data.“ In: *Proceedings of the 12th International Conference on Information Quality*. 2007 (zitiert auf S. 25, 49).

[Hün11] K. M. Hüner. „Führungssysteme und ausgewählte Massnahmen zur Steuerung von Konzerndatenqualität“. In: 2011 (zitiert auf S. 21).

[HVS+16] J. He, E. Veltri, D. Santoro, G. Li, G. Mecca, P. Papotti, N. Tang. „Interactive and deterministic data cleaning: A tossed stone raises a thousand ripples“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Bd. 26-June-2016. Association for Computing Machinery, Juni 2016. DOI: 10.1145/2882903.2915242. URL: http://dl.acm.org/citation.cfm?doid=2882903.2915242 (zitiert auf S. 27, 49).

[Int17] D. International. *DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition)*. Hrsg. von Intergovernmental Panel on Climate Change. Technics Publications, 2017, S. 644. URL: https://dl.acm.org/doi/book/10.5555/3165209 (zitiert auf S. 19).

[IPC15] S. Idreos, O. Papaemmanouil, S. Chaudhuri. „Overview of Data Exploration Techniques“. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*. Bd. 2015. ACM Press, Mai 2015, S. 277–281. DOI: 10.1145/2723372.2731084. URL: http://dl.acm.org/citation.cfm?doid=2723372.2731084 (zitiert auf S. 17).

[IW11] C. Imhoff, C. White. *Self-Service Business: Intelligence Empowering Users to Generate Insights*. Techn. Ber. tdwi, 2011. URL: http://triangleinformationmanagement.com/wp-content/uploads/2014/02/Self-Service-Business-Intelligence-empowering-users-to-generate-insights.pdf (zitiert auf S. 11).

[JGDW18] S. Juddoo, C. George, P. Duquenoy, D. Windridge. „Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context“. In: *Applied System Innovation* 1 (Nov. 2018). DOI: 10.3390/asi1040043. URL: http://www.mdpi.com/2571-5577/1/4/43 (zitiert auf S. 18).

[JPR19] A. Jadhav, D. Pramod, K. Ramanathan. „Comparison of Performance of Data Imputation Methods for Numeric Dataset“. In: *Applied Artificial Intelligence* (Juli 2019), S. 1–21. DOI: 10.1080/08839514.2019.1637138 (zitiert auf S. 51, 54).

[Jud15] S. Juddoo. „Overview of data quality challenges in the context of Big Data“. In: *2015 International Conference on Computing, Communication and Security (ICCCS)*. IEEE, Dez. 2015, S. 1–9. DOI: 10.1109/CCCS.2015.7374131. URL: http://ieeexplore.ieee.org/document/7374131/ (zitiert auf S. 24, 49).

[Kai10] M. Kaiser. „A conceptional approach to unify completeness, consistency, and accuracy as quality dimensions of data values“. In: *Proceedings of the European, Mediterranean and Middle Eastern Conference on Information Systems: Global Information Systems Challenges in Management, EMCIS 2010*. 2010, S. 17. URL: https://www.researchgate.net/publication/200751113 (zitiert auf S. 18, 19, 23, 51, 52).

[KFG+16] S. Krishnan, M. J. Franklin, K. Goldberg, J. Wang, E. Wu. „ActiveClean: An interactive data cleaning framework for modern machine learning“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Association for Computing Machinery, Juni 2016. DOI: 10.1145/2882903.2899409. URL: http://dl.acm.org/citation.cfm?doid=2882903.2899409 (zitiert auf S. 27).

[KMSZ09] D. A. Keim, F. Mansmann, A. Stoffel, H. Ziegler. „Visual Analytics“. In: *Encyclopedia of Database Systems*. Hrsg. von L. LIU, M. T. ÖZSU. Boston, MA: Springer US, 2009. DOI: 10.1007/978-0-387-39940-9_1122. URL: https://doi.org/10.1007/978-0-387-39940-9_1122 (zitiert auf S. 11).

[KPHH11] S. Kandel, A. Paepcke, J. Hellerstein, J. Heer. „Wrangler: Interactive visual specification of data transformation scripts“. In: *Conference on Human Factors in Computing Systems - Proceedings*. ACM Press, 2011, S. 3363–3372. DOI: 10.1145/1978942.1979444. URL: http://dl.acm.org/citation.cfm?doid=1978942.1979444 (zitiert auf S. 27).

[KTSJ14] K. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan. „A scalable bootstrap for massive data“. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76.4 (2014), S. 795–816. URL: http://www.jstor.org/stable/24774569 (zitiert auf S. 73, 74).

[KUG14] M. A. U. D. Khan, M. F. Uddin, N. Gupta. „Seven V's of Big Data understanding Big Data to extract value“. In: *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education - Ëngineering Education: Industry Involvement and Interdisciplinary Trends”, ASEE Zone 1 2014*. IEEE Computer Society, 2014. DOI: 10.1109/ASEEZone1.2014.6820689 (zitiert auf S. 17).

[Lan01] D. Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Techn. Ber. Gartner, Feb. 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (zitiert auf S. 17).

[LL19] C. Lennerholt, J. van Laere. „Data Access and Data Quality Challenges of Self-Service Business Intelligence“. In: *ECIS*. 2019 (zitiert auf S. 14).

[Los11] D. Loshin. *The Practitioner's Guide to Data Quality Improvement*. San Francisco, CA, United States: Morgan Kaufmann Publishers Inc.340 Pine Street, 2011, S. 327–350. URL: http://www.sciencedirect.com/science/article/pii/B9780123737175000191 (zitiert auf S. 21, 24, 48).

[LPFW06] Y. W. Lee, L. L. Pipino, J. D. Funk, R. Y. Wang. *Journey to Data Quality*. The MIT Press, 2006, S. 240 (zitiert auf S. 19, 20, 25, 54).

[LPW04] Y. W. Lee, L. Pipino, R. Y. Wang. „Process-Embedded Data Integrity“. In: 15 (2004), S. 87–103 (zitiert auf S. 20, 54).

[LSKW02] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang. „AIMQ: A methodology for information quality assessment“. In: *Information and Management* 40.2 (Dez. 2002), S. 133–146. DOI: 10.1016/S0378-7206(02)00043-5 (zitiert auf S. 32).

[MCR+16] J. Merino, I. Caballero, B. Rivas, M. Serrano, M. Piattini. „A Data Quality in Use model for Big Data“. In: *Future Generation Computer Systems* 63 (Okt. 2016), S. 123–130. DOI: 10.1016/j.future.2015.11.024. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X15003817 (zitiert auf S. 32).

[MR10] O. Maimon, L. Rokach. *Data Mining and Knowledge Discovery Handbook*. 2nd. Springer Publishing Company, Incorporated, 2010 (zitiert auf S. 11).

[MTV19] G. Mylavarapu, J. P. Thomas, K. A. Viswanathan. „An Automated Big Data Accuracy Assessment Tool“. In: *2019 4th IEEE International Conference on Big Data Analytics, ICBDA 2019*. Institute of Electrical und Electronics Engineers Inc., Mai 2019, S. 193–197. DOI: 10.1109/ICBDA.2019.8713218 (zitiert auf S. 48, 49).

[Nic07] T. G. Nick. „Descriptive Statistics“. In: *Topics in Biostatistics*. Hrsg. von W. T. Ambrosius. Totowa, NJ: Humana Press, 2007, S. 33–52. DOI: 10.1007/978-1-59745-530-5_3. URL: https://doi.org/10.1007/978-1-59745-530-5_3 (zitiert auf S. 44, 45, 58).

[OG08] C. Ordonez, J. García-García. „Referential integrity quality metrics“. In: *Decision Support Systems* 44 (Jan. 2008), S. 495–508. DOI: 10.1016/j.dss.2007.06.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167923607000887 (zitiert auf S. 20, 54).

[OMG19] O. M. G. (OMG). *Decision Model and Notation (DMN) Specification, Version 1.3*. OMG File ID dtc/19-12-06 (https://www.omg.org/spec/DMN/1.3/). 2019 (zitiert auf S. 48, 49).

[PLW02] L. L. Pipino, Y. W. Lee, R. Y. Wang. „Data quality assessment“. In: *Communications of the ACM* 45 (Apr. 2002). DOI: 10.1145/505248.506010. URL: http://portal.acm.org/citation.cfm?doid=505248.506010 (zitiert auf S. 20, 21).

[RCIR17] T. Rekatsinas, X. Chu, I. F. Ilyas, C. Ré. „HoloClean: Holistic Data Repairs with Probabilistic Inference“. In: 10 (Feb. 2017). eprint: 1702.00820. URL: http://arxiv.org/abs/1702.00820 (zitiert auf S. 27).

[RGR18] D. Reinsel, J. Gantz, J. Rydning. *The Digitization of the World From Edge to Core*. Techn. Ber. IDC, Nov. 2018. URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (zitiert auf S. 11).

[RWX+07] E. A. Rundensteiner, M. O. Ward, Z. Xie, Q. Cui, C. V. Wad, D. Yang, S. Huang. „XmdvtoolQ:: Quality-aware interactive data exploration“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. New York, New York, USA: ACM Press, 2007, S. 1109–1112. DOI: 10.1145/1247480.1247623. URL: http://portal.acm.org/citation.cfm?doid=1247480.1247623 (zitiert auf S. 27).

[SBG+19] S. Shrivastava, A. Bhamidipaty, W. Gifford, S. Siegel, V. Ganapavarapu, J. Kalagnanam. „DQA: Scalable, Automated and Interactive Data Quality Advisor“. In: *2019 IEEE International Conference on Big Data (Big Data)*. Institute of Electrical und Electronics Engineers Inc., Dez. 2019, S. 2913–2922. DOI: 10.1109/BigData47090.2019.9006187 (zitiert auf S. 27, 28, 32).

[SC02] M. Scannapieco, T. Catarci. „Data Quality under the Computer Science perspective“. In: *Journal of The ACM - JACM* 2.2 (2002), S. 1–12 (zitiert auf S. 18).

[SETN16] M. A. Serhani, H. T. El Kassabi, I. Taleb, A. Nujum. „An Hybrid Approach to Quality Evaluation across Big Data Value Chain“. In: *2016 IEEE International Congress on Big Data (BigData Congress)*. IEEE, Juni 2016, S. 418–425. DOI: 10.1109/BigDataCongress.2016.65. URL: http://ieeexplore.ieee.org/document/7584971/ (zitiert auf S. 24, 32).

[SMB05] M. Scannapieco, P. Missier, C. Batini. „Data Quality at a Glance“. In: *Datenbank-Spektrum* 14 (2005) (zitiert auf S. 18–20, 23).

[Ste46] S. S. Stevens. „On the Theory of Scales of Measurement“. In: *Science* 103 (Juni 1946), S. 677–680. DOI: 10.1126/science.103.2684.677. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.103.2684.677 (zitiert auf S. 21).

[TKS+16] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, C. Bouhaddioui. „Big Data Quality: A Quality Dimensions Evaluation“. In: *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. IEEE, Juli 2016, S. 759–765. DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122. URL: https://ieeexplore.ieee.org/document/7816918/ (zitiert auf S. 32, 33, 70, 73, 74).

[TKS04] P.-N. Tan, V. Kumar, J. Srivastava. „SSelecting the right objective measure for association analysis““. In: *Ïnformation Systems”* ”29” (”2004”). ”Knowledge Discovery and Data Mining (KDD 2002)”. DOI: "https://doi.org/10.1016/S0306-4379(03)00072-3". URL: http://www.sciencedirect.com/science/article/pii/S0306437903000723 (zitiert auf S. 49).

[VMH16] R. Vaziri, M. Mohsenzadeh, J. Habibi. „TBDQ: A Pragmatic Task-Based Method to Data Quality Assessment and Improvement“. In: *PLOS ONE* 11.5 (Mai 2016). Hrsg. von J. Hewitt, e0154508. DOI: 10.1371/journal.pone.0154508. URL: https://dx.plos.org/10.1371/journal.pone.0154508 (zitiert auf S. 32).

[VPV+19] Á. Valencia, L. Parody, A. Varela Vaca, I. Caballero, M. T. Gómez López. „DMN for Data Quality Measurement and Assessment“. In: *Business Process Management Workshops*. Jan. 2019, S. 362–374. DOI: 10.1007/978-3-030-37453-2_30 (zitiert auf S. 48, 49).

[WLG16] H. Wang, J. Li, H. Gao. „Data Inconsistency Evaluation for Cyberphysical System“. In: *International Journal of Distributed Sensor Networks* 12 (Aug. 2016). DOI: 10.1177/155014779496878. URL: http://journals.sagepub.com/doi/10.1177/155014779496878 (zitiert auf S. 25).

[WW96] Y. Wand, R. Y. Wang. „Anchoring data quality dimensions in ontological foundations“. In: *Communications of the ACM* 39 (1996). DOI: 10.1145/240455.240479 (zitiert auf S. 18, 20).

[YP10] P. Z. Yeh, C. A. Puri. „An Efficient and Robust Approach for Discovering Data Quality Rules“. In: *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*. Bd. 1. IEEE, Okt. 2010. DOI: 10.1109/ICTAI.2010.43. URL: http://ieeexplore.ieee.org/document/5670046/ (zitiert auf S. 49).

[Zmu78] R. W. Zmud. „AN EMPIRICAL INVESTIGATION OF THE DIMENSIONALITY OF THE CONCEPT OF INFORMATION“. In: *Decision Sciences* 9.2 (Apr. 1978), S. 187–195. DOI: 10.1111/j.1540-5915.1978.tb01378.x. URL: http://doi.wiley.com/10.1111/j.1540-5915.1978.tb01378.x (zitiert auf S. 18).

[ZZ16] N. Zellal, A. Zaouia. „A measurement model for factors influencing data quality in data warehouse“. In: *2016 4th IEEE International Colloquium in Information Science and Technology (CIST)*. Bd. 0. Institute of Electrical und Electronics Engineers Inc., Juli 2016, S. 46–51. DOI: 10.1109/CIST.2016.7805102 (zitiert auf S. 48).

Alle URLs wurden zuletzt am 25. 08. 2020 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift