Lehrstuhl für Wirtschafts- und Betriebswissenschaften

Dissertation

Reifegradmodell zur Bewertung der Inputfaktoren für datenanalytische Anwendungen-Konzeptionierung am Beispiel der Schwachstellenanalyse

Dipl.-Ing. Robert Bernerstätter, BSc

April 2019

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich diese Arbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt, und mich auch sonst keiner unerlaubten Hilfsmittel bedient habe.

Ich erkläre, dass ich die Richtlinien des Senats der Montanuniversität Leoben zu "Gute wissenschaftliche Praxis" gelesen, verstanden und befolgt habe.

Weiters erkläre ich, dass die elektronische und gedruckte Version der eingereichten wissenschaftlichen Abschlussarbeit formal und inhaltlich identisch sind.

Datum 26.04.2019

Unterschrift Verfasser/in
Robert, Bernerstätter
Matrikelnummer: 00735041

Gleichheitsgrundsatz

Aus Gründen der Lesbarkeit wurde in dieser Arbeit darauf verzichtet, geschlechtsspezifische Formulierungen zu verwenden. Es wird ausdrücklich festgehalten, dass die bei Personen verwendeten maskulinen Formen für beide Geschlechter zu verstehen sind.

Danksagung

Mein besonderer Dank gilt meinem Doktorvater o.Univ.-Prof. Dipl.-Ing. Dr. mont. Hubert Biedermann für die Betreuung der Arbeit. Seine Erfahrung in der wissenschaftlichen Vorgehensweise war eine wertvolle Hilfe bei der Erstellung dieser Arbeit. Darüber hinaus regten die fachlichen Diskussionen mit ihm die ständige kritische Reflexion und Verbesserung der Inhalte an. Außerdem ermöglichte er mir durch sein entgegen gebrachtes Vertrauen in meine Arbeit am Lehrstuhl eine unschätzbare Entwicklung meiner Persönlichkeit. Dafür Danke!

Bei Herrn Univ.-Prof. Dr.-Ing. Jochen Deuse will ich mich für die über Jahre hilfreichen Anregungen während meines Dissertationsfortschrittes bedanken.

Meine Kolleginnen und Kollegen am Lehrstuhl für Wirtschafts- und Betriebswissenschaften möchte ich nicht unerwähnt lassen. Sie haben durch ihre Inputs bei unseren zahlreichen wissenschaftlichen Austauschveranstaltungen einen wichtigen Beitrag zur Gestaltung dieser Arbeit geleistet.

Meiner Verlobten Andrea gilt ein besonderer Platz an dieser Stelle. Sie musste in den vergangenen Jahren immer wieder auf gemeinsame Aktivitäten verzichten. In den letzten intensiven Monaten der Entstehung dieser Arbeit war sie mir eine unverzichtbare Stütze. Vielen lieben Dank Schatz!

Kurzfassung

Die fortschreitende Durchdringung der Industrie mit digitalisierten und vernetzten Komponenten steigert in den letzten Jahren die erzeugte Datenmenge. Zur nutzbringenden Verwertung der Daten setzen immer mehr Unternehmen datenanalytische Projekte um. Dabei werden die Voraussetzungen, die für solche Vorhaben nötig sind, außer Acht gelassen oder falsch bewertet. Diese Voraussetzungen sind die Inputfaktoren des Datenanalyseprozesses, wie eine effektive Datenerfassung und eine effiziente Datenbereitstellung auf der einen Seite und der Inhalt der Daten in der klassischen Datenqualitätssicht auf der anderen Seite. Wenn diese beiden Facetten für ein Projekt nicht ausreichend entwickelt sind führt das nicht nur zu zeitlichen und folglich finanziellen Abweichungen in der Umsetzung, sondern im schlimmsten Fall zum Scheitern des Projektes und zu einem Ansehensverlust datenanalytischer Initiativen.

Innovative und zukunftsweisende Projekte auf dem Gebiet der Datenanalytik sollten verwendet werden, um die Strukturen im Unternehmen auf ein Niveau zu entwickeln und damit den kommenden Herausforderungen zu begegnen. Reifegradmodelle unterstützen bei der Bewertung von Unternehmensprozessen und der strukturierten Verbesserung selbiger. Unter den zahlreichen existierenden Reifegradmodellen der Digitalisierung gibt es jedoch keines, welches sich mit der fokussierten Bewertung des datenanalytischen Prozesses mit dem Schwerpunkt seiner benötigten Inputfaktoren, wie Datenmanagement und Datenqualität beschäftigt.

Das in dieser Arbeit entwickelte Reifegradmodell soll diese Lücke schließen. Der CRISP-DM als generisches Prozessmodell zur Datenanalyse liegt der Bewertung zu Grunde. Der strukturierte Aufbau berücksichtigt gängige Datenqualitätsdimensionen, um Bewertungsanforderungen auf die Ebene der Reifegradkategorien herunterzubrechen. Die sechs Reifegradkategorien sind so gewählt, dass in jeder von ihnen praxisorientierte Handlungsempfehlungen abgegeben werden können, um Verbesserung im Reifegrad zu erzielen. Die Hierarchie der Reifegradstufen orientiert sich an der steigenden Komplexität der Analysekonzepte, deren Einsatz einen verstärkten Unternehmensnutzen bringen.

Das Reifegradmodell wurde an Fallbeispielen entwickelt und getestet. In diesem Rahmen wurde eine Methode für eine Big Data gestützte Schwachstellenanalyse aus dem klassischen Methodenkoffer der Datenanalyse verwendet und die Ergebnisse in einem datengestützten Ishikawa Diagramm dargestellt.

Abstract

Due to the penetration of the industry with digitized and connected components, the amount of data has increased in recent years. More and more companies are implementing data analysis projects in order to be able to utilize this data profitably. The prerequisites necessary for such projects are either ignored or incorrectly evaluated. These prerequisites are the input factors of the data analysis process, such as effective data acquisition and efficient data provision on the one hand and the content of the data in the classical data quality view on the other. If these two facets are not sufficiently developed for a project, it leads not only to time and consequently financial deviations in the implementation, but in the worst case to the failure of the project and to a loss of reputation of data analysis initiatives.

Innovative and forward-looking projects in the field of data analytics should be used to develop structures in the company to a level where they can meet future challenges. Maturity models support the evaluation of business processes and their structured improvement. Among numerous existing maturity models for digitization, however, there is none that deals with the focused evaluation of the data analytical process with a focus on its required input factors, such as data management and data quality.

The maturity model developed in this thesis should close this gap. The evaluation is based on the CRISP-DM as a generic process model for data analysis. The structure considers common data quality dimensions in order to break down assessment requirements to the level of maturity categories. Six maturity level categories are selected in such a way that practice-oriented recommendations for action can be made in each of them in order to achieve an improvement in the maturity level. The hierarchy of maturity levels is oriented towards the increasing complexity of analysis concepts, the use of which increases company benefits. The maturity model was developed and tested using case studies. In this context a method for a Big Data supported weak point analysis from the classical method case of data analysis was used and the results were presented in a data supported Ishikawa diagram.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BDE	Betriebsdatenerfassung
BI	Business Intelligence
CIO	Cief Information Officer
CMM	Capability Maturity Model
CMMI	Capability Maturity Model Integration
CRISP-DM	Cross Industry Standard Process Data Mining
DGIQ	Deutsche Gesellschaft für Informationsqualität
DIN	Deutsches Institut für Normung
DMAIC	Define-Measure-Analyse-Improve-Control
DMKD	Data Mining and Knowledge Discovery
DSR	Design Science Research
ERP	Enterprise Ressource Planung
ETL	Extract, Transform and Load
FP	Frequent Pattern
FTM	Fixed Time Maintenance
ggf.	Gegebenenfalls
HTML	Hypertext Markup Language
ID	Identifikation
IEC	International Electrotechnical Organisation
IKT	Informations- und Kommunikatonstechnologie
ISO	International Organisation for Standardization
ISR	Information System Research
IT	Informationstechnologie
KDD	Knowledge Discovery in Databases
KDID	Knowledge Discovery in Industrial Databases
KVP	Kontinuierlicher Verbesserungsprozess
LSM	Lean Smart Maintenance
M2M	Machine to Machine (Maschine zu Maschine)
MDE	Maschinendatenerfassung

MES	Management Execution System
ms	Milisekunde(n)
OLAP	Online Analytical Processing
OPC-UA	Open Platform Communications Unified Architecture
OT	Operative Technologie
PA	Prozessattribut
PDF	Portable Document Format
SCM	Supply Chain Management
SEI	Software Engineering Institute
SEMMA	Sample, Explore, Modify, Model, Assess
SPICE	Software Process Improvement and Capability Determination
SPS	Speicherprogrammierte Steuerung
SQL	Structured Query Language
TDQM	Total Data Quality Management
tlw	teilweise
u. a.	unter anderen
u. U.	unter Umständen
Vgl.	vergleiche
XML	Extensible Markup Language

1 Einleitung

Dieses Einleitungskapitel beschreibt die Ausgangssituation und Problemstellung, welche zu dieser Arbeit motiviert haben und bildet in diesem Sinne die Rechtfertigung für ihre Erstellung. Es wird folgend auf die Zielsetzung der Dissertation und den daraus abgeleiteten Forschungsfragen eingegangen. Abschließend wir das Forschungsdesign, dem diese Arbeit zugrunde liegt und der generelle Aufbau der Arbeit beschrieben.

1.1 Ausgangssituation und Problemstellung

Die Digitalisierung in der Industrie fand mit der Einführung des Begriffs Industrie 4.0 im Jahr 2011 auf der Hannover Messe[1] einen Ausdruck, der die Entwicklungen auf diesem Gebiet zusammenfasst. Zwei Jahre später wurde von den Initiatoren KAGERMANN, LUKAS und WAHLSTER als Mitglieder der Promotorengruppe Kommunikation der Forschungsunion Wirtschaft-Wissenschaft der Abschlussbericht des Arbeitskreises Industrie 4.0 mit dem Titel „Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0“ vorgelegt. Einer der wesentlichen Befähiger und Treiber von Industrie 4.0 ist die durchgehende Vernetzung entlang der gesamten Wertschöpfungskette und über die Hierarchiegrenzen hinweg.[2]

Die verstärkte Vernetzung der Systeme konfrontiert Unternehmen und deren Mitarbeiter mit einer zunehmenden Komplexität die schwer zu verstehen und beherrschen ist.[3] Sie versuchen diese Komplexität durch komplexe Analysemethoden, die durch die Digitalisierung weiter verbreitet werden, zu beherrschen. Die wachsende Menge an Daten soll so nützlich verwendet werden. Die Anwendung in der Analyse von Fehlersymptomen und –ursachen im Störungsmanagement ist daher ein stark wachsendes Tätigkeitsfeld von Big Data Möglichkeiten wie Data-Mining.[4]

Die Intensivierung der Kommunikation der unterschiedlichen Systemkomponenten führt zu einem Anstieg der erzeugten Daten. Diese ermöglichen neben innovativen Geschäfts- und Servicemodellen[5] auch neue Einblicke in den eigenen Prozess zur Hebung von Effizienzpotenzialen.[6] Zur Nutzung der Möglichkeiten setzen immer mehr Unternehmen auf den Einsatz von Data-Mining und den damit verbundenen Methoden. Umfragen zeigen, dass ca. 70% der Unternehmen keine Maßnahmen im Bereich Data-Mining umgesetzt haben, bzw. in den nächsten 3-5 Jahren Data-Mining anwenden

[1] Vgl. VDI Verlag, https://www.vdi-nachrichten.com/Technik-Gesellschaft/Industrie-40-Mit-Internet-Dinge-Weg-4-industriellen-Revolution (Zugriff: 14.03.2019)
[2] Vgl. Promotorengruppe Kommunikation der Forschungsunion Wirtschaft – Wissenschaft (2013), S. 6.
[3] Vgl. Budde, L.; Friedli, T. (2017), S. 29 ff.
[4] Vgl. Bange, C.; Janoschek, N. (2014), S. 33.; Schmitz, S.; Krenge, J. (2014), S. 1299 f.
[5] Vgl. Promotorengruppe Kommunikation der Forschungsunion Wirtschaft – Wissenschaft (2013), S. 20.
[6] Vgl. Ketteler, D.; König, C. (2017), S. 7.

wollen.[7] Nur ein Drittel der Unternehmen sind zufrieden mit den Ergebnissen aus datenanalytischen Projektanwendungen.[8]

Neben der fehlenden Qualifikation und unzureichender Erfahrung der Mitarbeiter in den Unternehmen, sind auch infrastrukturelle und systemische Ursachen, wie die Kompatibilität zwischen den IT-Systemen und Datenquellen, sowie eine mangelnde Qualität der Daten als Gründe für die Probleme angeführt.[9] Sie führen dazu, dass der Anwendungsfall unzureichend abgegrenzt und den Möglichkeiten des Unternehmens angepasst wird.

Die Inkompatibilität und die mangelnde Datenqualität sind die Hauptgründe für eine zeitaufwendige oder gar unmögliche Datenaufbereitung.[10] Zahlreiche Studien zeigen, dass speziell die vorbereitenden Maßnahmen der Analyse, wie die Datenaufbereitung, und nicht die Analyse selbst oder der Implementierungsprozess der Ergebnisse, die meiste Zeit in datenanalytischen Projekten in Anspruch nehmen. Tabelle 1 zeigt eine Übersicht der unterschiedlichen Werte der Zeitverbräuche in Data-Mining Projekten. Bei CIOS & KURGAN und MUNSON handelt es sich um die jeweiligen Medianwerte, weshalb sich die Summe nicht zwingend auf 100% ergänzt.

Die Zeit, die in die Datenaufbereitung investiert wird, ist jedoch nötig, da nur mit einer hohen Datenqualität es den Unternehmen möglich ist, die nötigen Erkenntnisse durch den Analyseprozess zu gewinnen, der ihnen einen Wettbewerbsvorteil verschafft[11].

Tabelle 1: Zeitaufwände in Data-Mining Projekten[12]

Prozessschritte	HEIMES ET AT.	CIOS & KURGAN	MUNSON
Geschäftsmodell verstehen	11,2%	15%	10%
Daten verstehen	14,5%	15%	Keine Angabe (nicht herauslesbar)
Daten aufbereiten	33,7% (inkl. Daten sammeln)	45%	50% (inkl. Daten sammeln)
Modellierung (Data-Mining)	14,0%	18%	18%
Evaluation	13,7	8%	11%
Einsatz	12,9%	8%	11%

Die Datenqualität war auch in der Vergangenheit ein wichtiger Faktor in Unternehmen. Die Sicht beschränkte sich jedoch auf die Richtigkeit von Berichten und die Effizienz diese zu erstellen. Durch die Digitalisierung wird die Datenqualität jedoch ein entscheidender Wettbewerbsfaktor. Daten werden in diesem Zusammenhang immer

[7] Vgl. Bange, C.; Janoschek, N. (2014), S. 17; Lueth, K. L. et al. (2016), S. 17.
[8] Vgl. Lueth, K. L. et al. (2016), S. 23 f.
[9] Vgl. Bange, C.; Janoschek, N. (2014), S. 23 f.; Lorenz, M. et al. (2016), S. 6 f.; Lueth, K. L. et al. (2016), S. 49.
[10] Vgl. Stodder, D. (2016), S. 19 f.
[11] Vgl. Klier, M.; Heinrich, B. (2016), S. 488.
[12] Quelle: Eigene Darstellung in Anlehnung an Cios, K. J.; Kurgan, Lukasz A. (2004), S. 10.; Munson, M. A. (2012), S. 67.; Heimes, H. et al. (2019), S. 60.

stärker als zusätzlicher Produktionsfaktor betrachtet, dessen Qualität es ebenso rigoros zu überwachen gilt wie die eines Rohstoffes in der Produktion.[13]

In der Vergangenheit wurden für die Messung der Datenqualität Systeme geschaffen, die sich verstärkt auf die Stammdaten konzentrierten.[14] Diese statischen Ansätze sind für die Datenlandschaft von Industrie 4.0 die sich durch Bewegungsdaten auszeichnet nicht geeignet. In diesem Zusammenhang sind der Datenzugang, die Datenaufbereitung und der Analyseprozess die Hauptkostentreiber, Bereiche, die in der klassischen Datenqualitätssicht nicht betrachtet werden.[15]

Da die Analysen im Rahmen von Data-Mining regelmäßig und in einigen Anwendungen in Echtzeit durchgeführt werden, müssen die Daten nicht nur von hoher Qualität sein. Es ist erforderlich, dass auch die Datenerfassung und –bereitstellung eine hohe Reife besitzen.[16] Die Unternehmen müssen in der Lage sein den datenanalytischen Prozess des Data-Mining so durchzuführen, dass die unterschiedlichen Aufgabenstellungen gelöst werden. Das Zusammenspiel von Analyseprozess, dem System und der Anwendung ist dabei wesentlich.[17] Um Prozesse und Betriebe zu bewerten eigenen sich Reifegradmodelle. In der Literatur gibt es jedoch kein Reifegradmodell, welches die Fähigkeit von Organisationen feststellt, Data-Mining durchzuführen und sich dabei auf die Daten und deren Generierung stützt.

1.2 Zielsetzung und Forschungsfrage

Basierend auf der beschriebenen Ausganssituation ist es das Ziel dieser Arbeit ein Reifegradmodell zu entwickeln, welches bewertet ob die Daten in ihrer Bereitstellung und in ihrem Inhalt geeignet sind unterschiedlich komplexe Analysen zu erlauben. Es soll dabei helfen den Anwendungsfall besser abzugrenzen und die Möglichkeiten zu erheben, bevor mit der Analysearbeit begonnen wird und ggf. unnötig viel Zeit investiert wird, ohne dass Aussicht auf Erfolg besteht. Durch die Darstellung der Möglichkeiten und dem Abgleich der geforderten Aufgabenstellung sollen Handlungsempfehlungen gegeben werden, um die Reife bei Bedarf zu verbessern. Dadurch soll das Unternehmen befähigt werden die Analyse nach der Umsetzung der Empfehlungen durchzuführen. Um konkret umsetzbare Handlungsempfehlungen geben zu können müssen die Reifegradkategorien Bereiche bewerten, die die Bereitstellung und den Inhalt der Daten abbilden.

Das Reifegradmodell wird anhand der Schwachstellenanalyse entwickelt, da diese Unternehmen bei der Beherrschung der Komplexität unterstützt. Trotzdem muss der Anspruch bestehen, dieses auch für andere Analysen einzusetzen zu können. Um diesen Zweck zu erfüllen, muss sich das Modell an einem generischen Analyseprozess orientieren.

[13] Vgl. Hazen, B. T. et al. (2014), S. 72 f.
[14] Vgl. Klier, M.; Heinrich, B. (2016), S. 489.
[15] Vgl. Lueth, K. L. et al. (2016), S. 42.
[16] Vgl. Lueth, K. L. et al. (2016), S. 49.
[17] Vgl. Markl, V. et al. (2013), S. 8 ff.

Es ergibt sich daher folgende wissenschaftliche Hauptforschungsfrage:

„Wie muss ein Reifegradmodell aufgebaut und inhaltlich ausgestaltet sein, um die Faktoren bewerten zu können, welche die Datenqualität für die unterschiedlichen Komplexitäten des Data-Minings beeinflussen?“

Folgende Subfragen werden abgeleitet:

1. Mit welchen Reifegradkategorien und -stufen lassen sich diese Faktoren und Komplexitätsgrade beschreiben?
2. Lassen sich bestehende Datenqualitätsbetrachtungen in die Bewertungslogik integrieren bzw. was muss dafür verändert werden?
3. Sind Zusammenhänge zwischen der Reife des Datenmanagements und der Datenqualität feststellbar?
4. Wie kann eine Schwachstellenanalyse in die Logik der Analysekomplexität des Data-Minings integriert werden?
5. Wie müssen Reifegrade für die Schwachstellenanalyse beschrieben und wie kann der Reifegrad für diese erhoben werden?
6. Welche datenanalytischen Methoden können die Schwachstellenanalyse bei steigender Systemkomplexität unterstützen?

1.3 Forschungsdesign

Zur Konkretisierung der vorstehenden Forschungsfragen wurde nach dem Design Science Research (DSR) Ansatz von HEVNER[18] vorgegangen, welcher aus dem Information System Research (ISR) Framework hervorgegangen ist.[19] Im Design Science wird zusätzliche Erkenntnis in Form von Konstrukten gewonnen. Das können Techniken, Methoden oder Modelle sein.[20] Design Research erforscht Design an sich und das Verständnis von Design. DSR vereint Design Science und Design Research. Es lernt und erweitert den wissenschaftlichen Horizont durch die Erzeugung von Artefakten. Dieses Lernen durch die laufende Anwendung in der Artefaktentwicklung ist die ausschlaggebende Weiterentwicklung von DSR gegenüber Design Science und Design Research.[21]

Das Forschungsframework von DSR basiert auf jenem von ISR. Abbildung 1 zeigt die Zyklen, die in Design Science Research angewandt werden. Sie greifen in der Entwicklung der Designartefakte ineinander und gewinnen dadurch die Legitimation des Forschungsvorhabens aus der Anwendungsdomäne der Umwelt und der Fundierung aus der Literatur. Sie tragen durch die Erzeugung des Artefakts zum praktischen Umfeld und durch wissenschaftliche Erkenntnisse und Veröffentlichungen zur Wissensbasis auf diese Weise zu den beiden Feldern bei.[22]

[18] Vgl. Hevner, A. R. (2007)
[19] Vgl. Hevner, A. R. et al. (2004)
[20] Vgl. Vaishnavi, V.; Kuechler, W. (2015), S. 11.
[21] Vgl. Vaishnavi, V.; Kuechler, W. (2015), S. 13.
[22] Vgl. Cronholm, S.; Göbel, H. (2016), S. 162.

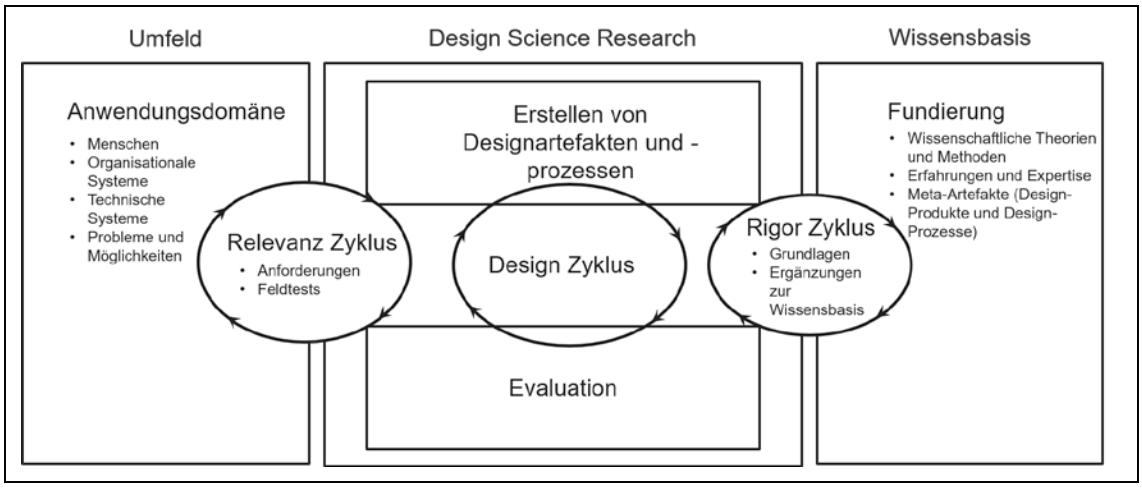

Abbildung 1: Design Science Research Zyklen[23]

Der DSR Ansatz hat zum Ziel das praktische Umfeld durch die neuen Artefakte zu entwickeln. Der Relevanz Zyklus stößt ein Forschungsvorhaben im Rahmen des DSR an indem Anforderungen aus dem Umfeld aufgegriffen werden. Dabei werden zusätzlich die Eingangsgrößen definiert und das System, für welches das Artefakt seine Gültigkeit hat, abgegrenzt.[24]

Auslöser dieser Arbeit waren Datenanalytikprojekte, die aufgrund unrealistischer Einschätzungen unbefriedigende Ergebnisse lieferten oder den vorgegebenen Zeitrahmen deutlich überschritten. Die Gründe waren, dass unvorhergesehene Begleitmaßnahmen durchgeführt werden mussten, bevor die Daten in ausreichender Qualität vorlagen. Problematisch waren somit die Datenqualität und in dieser Verbindung die Systeme zur Datenaufnahme und -bereitstellung. Die Bewertungsmethoden der Literatur waren auf den industriellen Bereich nicht übertragbar wodurch sich die Forschungslücke ergab. Die Systemgrenze bildet ein Betrieb und anlagenrelevante Ausfälle und Schwachstellen. Betrachtet werden die Daten, die innerhalb dieses Systems erzeugt, ausgetauscht und in letzter Konsequenz aufgezeichnet werden. Qualitative Beurteilungen werden nicht berücksichtigt, da als Teil der Arbeit eine Methode gefunden werden soll, die sich für die datengestützte Schwachstellenanalyse eignet. Rückmeldungen von Folgeprozessschritten einer identifizierten Schwachstelle werden nicht beachtet. Des Weiteren wird nicht der Anspruch erhoben organisatorische Schwachstellen zu untersuchen, da für diese die Datenbasis in den Fallbeispielen nicht ausreichend ist.

Menschliche Schwachstellen oder der menschliche Einfluss als Ursache für Schwachstellen sind ausgenommen, da dieser Einfluss aufgrund der weitreichenden Palette an Einflussfaktoren nicht quantitativ und deterministisch abbildbar ist. So wird die Qualifikation von Personen unterschiedlich zum Einsatz gebracht.[25]

Der Rigor Zyklus bindet die vorhandene Basis wissenschaftlicher Theorien in das Forschungsvorhaben ein. Dazu zählt der Stand der Technik von Methoden, existierende Artefakte in Form von bereits entwickelten Prozessen und die Erkenntnisse bereits

[23] Quelle: Eigene Darstellung aus dem Englischen übersetzt nach Hevner, A. R. (2007), S. 88.
[24] Vgl. Hevner, A. R. (2007), S. 89.
[25] Vgl. Heil, M. (1995), S. 88 f.

abgeschlossener Projekte. Durch die Berücksichtigung dieser Basis wird gewährleistet, dass die neu entwickelten Artefakte innovativ sind.[26]

Der Design Zyklus ist der Kernbereich eines jeden Design Science Research Projektes. Hier wird das zentrale Designartefakt entwickelt, getestet und die Rückmeldungen und Ergebnisse des Tests fließen in die Weiterentwicklung ein. Dieses Vorgehen wiederholt sich, bis das Artefakt einen zufriedenstellenden Entwicklungsstatus erreicht hat.[27]

Das zentrale Artefakt dieser Arbeit ist das Reifegradmodell. Durch die prozessnahe Entwicklung dieser Modelle eignet sich der DSR Ansatz sehr gut. Die laufende Anpassung an unterschiedliche Fallbeispiele garantiert die allgemeine Gültigkeit über die Anwendungsbeispiele hinaus.[28] Die Entwicklung und die Evaluation die im Rahmen des Design Zyklus stattfanden sind in den Kapiteln 6 und 7 beschrieben.

1.4 Aufbau der Arbeit

Die Beantwortung der Forschungsfragen erfolgte im Rahmen des in Abbildung 2 dargestellten Aufbaus der Arbeit. Die acht Kapitel gliedern sich in die theoretische Fundierung, von Kapitel 2 bis 5, auf welche die empirische Umsetzung und Anwendung (Kapitel 6 und 7) aufbaut.

Kapitel 1 beschreibt die Ausgangssituation und die Problemstellung, die der Arbeit zu Grunde liegen. Die identifizierte Forschungslücke führt zur Zielsetzung der Arbeit, die durch die abgeleiteten Forschungsfragen geschlossen werden soll. Des Weiteren werden die Forschungsmethodik und der abgrenzende Rahmen, für den die Arbeit Gültigkeit hat, beschrieben.

Kapitel 2 behandelt die relevante Modelltheorie und beschreibt in weiterer Folge Reifegradmodelle im Detail. Es wird ein Auszug wichtiger Reifegradmodelle der Digitalisierung vorgestellt und kritisch betrachtet, um die Forschungsrelevanz nochmals zu untermauern.

Mit der Betrachtung des Daten- und Informationsmanagements bildet Kapitel 3 einen wesentlichen Abschnitt dieser Arbeit. Neben der Abgrenzung der Begriffe Daten und Informationen sowie deren Management, wird die zentrale Thematik der Daten- und Informationsqualität aufbereitet.

Kapitel 4 erörtert die fundamentalsten Grundlagen zur Datenanalyse. Darin werden die unterschiedlichen generischen Analyseprozesse vorgestellt aus denen ein geeigneter gewählt wird, der als Grundlage für das Reifegradmodell dient. Außerdem werden die verschiedenen Analysemethoden beschrieben, um eine für die Unterstützung der Schwachstellenanalyse passende zu wählen.

Die Schwachstellenanalyse wird in Folge in Kapitel 5 beschrieben. Wichtig ist dabei die Generalisierung und Abgrenzung des Begriffs von der singulär fokussierten Schadensanalyse in der Instandhaltung. Der Anspruch der Arbeit besteht in der

[26] Vgl. Hevner, A. R. (2007), S. 89 f.
[27] Vgl. Hevner, A. R. et al. (2004), S. 90.
[28] Vgl. Becker, J. et al. (2009), S. 250.

Ableitung einer allgemein gültigen Definition des Begriffs Schwachstelle, der den Komplexitätsanforderungen von hochvernetzten Produktionssystemen gerecht wird.

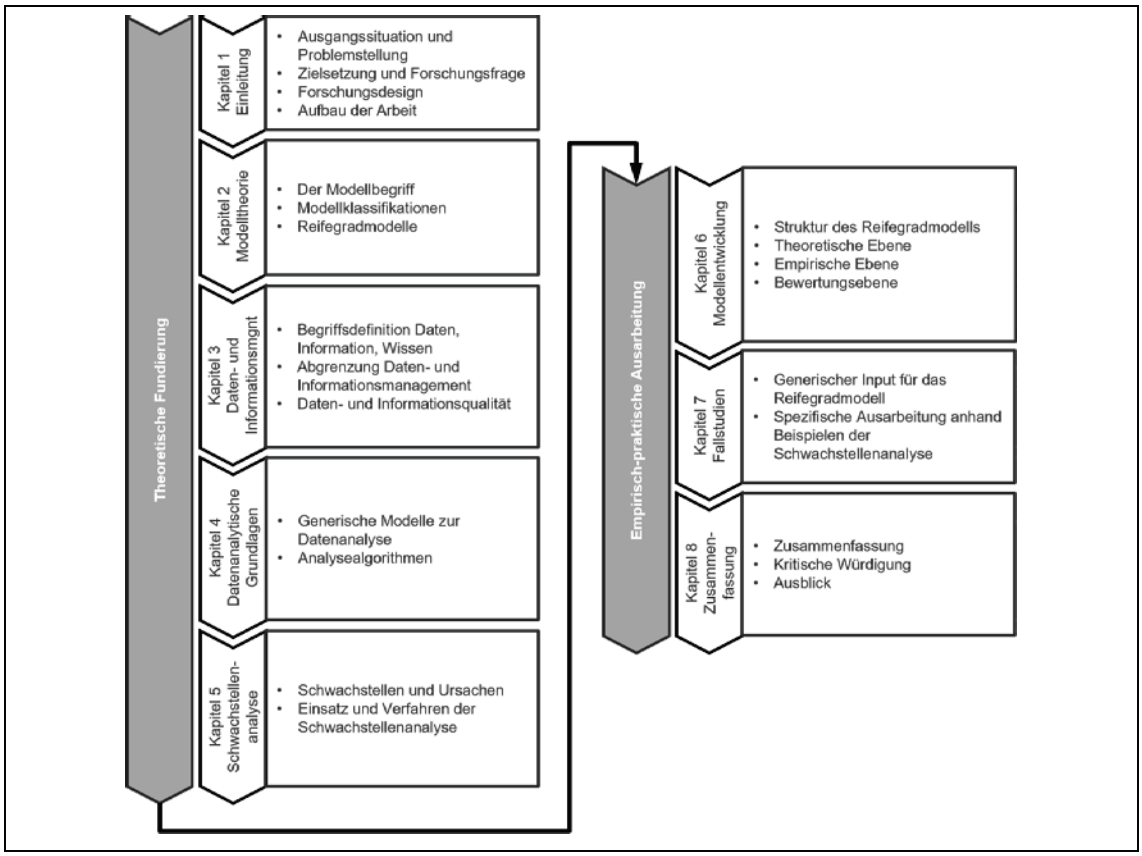

Abbildung 2: Aufbau der Arbeit[29]

Das zentrale Artefakt der Arbeit, das Reifegradmodell, bildet den Inhalt von Kapitel 6. Es wird der Aufbau und die Struktur erörtert. Dazu zählen die Reifegradkategorien und -stufen, deren logischer Aufbau aufeinander und das Zusammenspiel mit dem gewählten datenanalytischen Kernprozess und den in Kapitel 3 beschriebenen und teilweise adaptierten Datenqualitätsdimensionen.

Kapitel 7 zeigt den für den DSR-Prozess wichtigen stufenweisen Entwicklungsprozess des Reifegradmodells anhand von sechs Fallbeispielen. Dabei wurde es an drei generischen Beispielen entwickelt, um den Anspruch auf die universelle Anwendbarkeit zu erhalten. Die weiteren drei Fallbeispiele dienten der spezifischen Ausgestaltung anhand der Schwachstellenanalyse und der exemplarischen Anwendung des gewählten Analyseprozesses und der neu entwickelten Darstellung der Ergebnisse der Big Data gestützten Schwachstellenanalyse.

Kapitel 8 schließt mit einer Zusammenfassung der wesentlichen Erkenntnisse und der fokussierten Beantwortung der Forschungsfragen. Des Weiteren werden die Ergebnisse kritisch gewürdigt und daraus der weitere sich aus dieser Arbeit ergebende Forschungsbedarf aufgezeigt.

[29] Quelle: Eigene Darstellung

2 Modelltheorie

Das zentrale Artefakt dieser Arbeit ist ein Reifegradmodell. Diese erfordert es sich mit den Grundlagen des Modellbegriffs zu beschäftigen. In weiterer Folge werden Reifegradmodelle in eine allgemeine Klassifikationsstruktur eingeordnet. Es werden deren Entstehung, die wesentlichen Inhalte und Möglichkeiten sie zu gestalten beschrieben. Am Ende werden grundlegende Reifegradmodelle, auf denen viele der neu entstandenen basieren, genauer analysiert und es wird ein Auszug wichtiger und aktueller Reifegradmodelle vorgestellt, die sich mit Digitalisierung und Daten beschäftigen. Die wichtigsten Inhalte werden zusammenfassend dargestellt und kritisch auf ihre Eignung zur Bewertung des datenanalytischen Prozesses betrachtet.

2.1 Der Modellbegriff

Modelle sind eine Abbildung der Realität. Zur Erzeugung der Abbildung wird ein Abstraktionsprozess durchlaufen, der die wesentlichen Merkmale des zu beschreibenden Bereiches erfasst, wodurch Modelle eine Vereinfachung der Wirklichkeit darstellen.[30] Für STACHOWIAK ist der Abbildungsprozess und folglich die Bildung von Modellen die Grundlage für den wissenschaftlichen Erkenntnisgewinn.[31] Der Abstraktionsprozess bringt es mit sich, dass die Abbildung unter dem Gesichtspunkt des Modellbenutzers (Subjektes) für eine bestimmte Zeitdauer und unter weiteren Einschränkungen für spezielle Objekte gilt. Das Modell setzt sich allgemein betrachtet aus einer finiten Anzahl von Attributen zusammen, die in einer formellen oder informellen Sprache beschrieben werden, die als Abkürzungen formuliert werden können und umfassendere Fachausdrücke und Gebiete beinhalten können.[32] Modelle bilden somit die Realität ab wobei diese Abbildung nicht allumfassend sein kann und sich diese an bestimmte Anforderungen orientiert.

Diese Eigenschaften haben Modelle unterschiedlicher Fachrichtungen gemein und werden als Abbildungsmerkmal, Verkürzungsmerkmal und pragmatisches Merkmal bezeichnet und folgend genauer definiert:[33]

- Abbildungsmerkmal: „Modelle sind stets Modelle von etwas, nämlich Abbildungen, Repräsentationen natürlicher oder künstlicher Originale, die selbst wieder Modelle sein können.“[34] Der Ursprung der Originale, ob natürlich oder technisch, symbolisch oder begrifflich, ist dabei nebensächlich. Original und Modell werden als Attributklassen definiert, wobei die Abbildung eine Zuordnung von Modellattributen zu Originalattributen entspricht.

[30] Vgl. Ninck, A. et al. (1998), S. 29.
[31] Vgl. Stachowiak, H. (1980), S. 53.
[32] Vgl. Stachowiak, H. (1992), S. 219 f.
[33] Vgl. Stachowiak, H. (1973), S. 131 ff.; Vgl. Töllner, A. et al. (2010), S. 8 f.
[34] Stachowiak, H. (1973), S. 131.

- Verkürzungsmerkmal: „Modelle erfassen im Allgemeinen nicht alle Attribute des durch sie repräsentierten Originals, sondern nur solche, die den jeweiligen Modellerschaffern und/oder Modellbenutzern relevant scheinen.“[35] Die Modellerstellung erfolgt unter der Annahme des Wissens aller Original- und Modellattribute. Dadurch ist es möglich zu bestimmen welche von Relevanz sind und durch das Modell abgebildet werden und welche irrelevant sind und vernachlässigt werden sollen. Die Bestimmung von Relevanz und Irrelevanz für ein Modell orientiert sich am Zweck des Modells (siehe hierzu das pragmatische Merkmal).
- Pragmatisches Merkmal: „Modelle sind ihren Originalen nicht per se eindeutig zugeordnet. Sie erfüllen ihre Ersetzungsfunktion a) für bestimmte - erkennende und/oder handelnde, modellbenutzende - Subjekte, b) innerhalb bestimmter Zeitintervalle und c) unter Einschränkung auf bestimmte gedankliche oder tatsächliche Operationen.“[36] Modelle bilden somit nicht nur etwas ab, sondern übernehmen diese Funktion für jemanden, für eine bestimmte Zeit und eine bestimmte Zielsetzung. Benützt sollen sie von Personen werden, die an der Entwicklung beteiligt waren oder über die Kompetenz verfügen sie anzuwenden.

Diese Eigenschaften sollten allen Modellen inhärent sein. Folgend wird die Einteilung der weitverbreiteten betriebswirtschaftlichen Modelle beschrieben. Diese Einteilung bietet gut geeignete Klassen, um auch Reifegradmodelle einzuteilen. Nichtsdestotrotz werden alternative Modellklassifizierungen behandelt, die über die betriebswirtschaftliche Sichtweise hinausgehen.

2.2 Modellklassifizierungen

Abhängig von der Forschungsdisziplin können Modelle unterschiedlichen Klassifikationen[37] zugeteilt werden. Hintergrund dieser Heterogenität sind die diversen Fragestellungen der Forschungsgebiete, die eine spezielle Sichtweise erfordern.

STACHOWIAK versucht eine möglichst generische und fachgebietsunabhängige Einteilung vorzunehmen, merkt jedoch an, dass diese der Eindeutigkeit und Ausschließlichkeit nicht genügt. Er unterscheidet in graphische Modelle, technische Modelle, semantische Modelle und semantisch-szientifische Modelle.[38]

Auch TROITZSCH strebt eine generische Klassifikation an. Er unterscheidet in der Modellbildung das Urbild, als das abzubildende Original, und das Bild, als das abgebildete Modell. Modellbildung heißt in diesem Fall den Urbildbereich in den Bildbereich abzubilden.[39] Je nach Bildbereich wird die Modelleinteilung vorgenommen.

[35] Stachowiak, H. (1973), S. 132.
[36] Stachowiak, H. (1973), S. 132 f.
[37] „Klassifikation bedeutet Klassen oder Gruppenbildung mit dem Ziel der Zuordnung von Objekten anhand bestimmter Begriffe, Regeln oder Kriterien „…“. Der Zweck von Klassifikationen ist es, durch Einteilung und Abgrenzung Ordnung in eine bestimmte Menge von Objekten zu bringen.“ Jockisch, M.; Rosendahl, J. (2010), S. 25.
[38] Vgl. Stachowiak, H. (1992), S. 220 ff.
[39] Vgl Troitzsch, K. G. (1990), S. 10.

In seinem Klassifikationsschema gibt es Realmodelle, ikonische Modell, Verbalmodell und Formalmodelle.[40]

2.2.1 Betriebswirtschaftliche Modelle

Modelle sind in der Betriebswirtschaft die wichtigste Möglichkeit, um Wissen und Erkenntnisse darzustellen.[41] Modelle werden dabei vordergründlich nach ihrer Zielsetzung unterschieden. Die Beschreibung der Attribute und deren Beziehungen, das Verständnis des Problems und die Lösung des Problems sind die drei essentiellen Ziele.[42] Tabelle 2 fasst die betriebswirtschaftlichen Modelle zusammen, die in weiterer Folge genauer beschrieben werden.

Tabelle 2: Klassifikation betriebswirtschaftlicher Modelle[43]

Zielsetzung	Modellklasse	Merkmale
Geordnete Beschreibung von Elementen und ihren Beziehungen in realen Systemen	Beschreibungsmodelle (Ermittlungs- und Erfassungsmodelle)	• Formulieren keine Hypothesen • Transformieren Daten in eine verständliche Form • Verwenden Definitionsgleichungen die auf einfachen arithmetischen Operationen beruhen
Beitrag zum Verstehen eines Problems	Erklärungs- oder Kausalmodelle	• Formulieren Hypothesen über Gesetzmäßigkeiten in realen Systemen • Beanspruchen empirische Geltung der gemachten Aussagen
	Prognosemodelle	• Sind dynamisch • Sagen zukünftige Entwicklungen voraus
Beitrag zum Lösen eines Problems	Entscheidungsmodelle	• Bewerten Handlungsalternativen im Rahmen einer Entscheidungssituation • Ermitteln Handlungsalternativen die einer gewissen Optimierungsbedingung genügen müssen
	Simulationsmodelle	• Untersuchen Wirkungsbeziehungen mit stochastischen Einflüssen • Konsequenzen von Handlungsalternativen werden auf ihre Zielwirkung untersucht

Beschreibungsmodelle

Sie werden auch deskriptive Modelle genannt und beschreiben Systeme und deren Elemente und die Beziehungen untereinander. Da keine Hypothesen über die Ursache-Wirkungszusammenhänge formuliert werden, sind eine Erklärung der Vorgänge sowie die Prognose selbiger nicht möglich.[44]

[40] Vgl. Troitzsch, K. G. (1990), S. 12 f.
[41] Vgl. Schweitzer, M. (1994), S. 53.
[42] Vgl. Schweitzer, M.; Küpper, H.-U. (1997), S. 5.
[43] Quelle: Homburg, C. (1998), S. 34. ergänzt durch Domschke, W.; Scholl, A. (2005), S. 31.
[44] Vgl. Homburg, C. (1998), S. 34.; Domschke, W.; Scholl, A. (2005), S. 31.

Erklärungs- oder Kausalmodelle

ADAM geht so weit und behauptet, dass alle Modelle wenigstens Erklärungsmodelle sind, da Modelle dem Abbildungsmerkmal folgen und die Struktur eines Systems erhalten müssen.[45] Erklärungsmodelle haben zum Ziel ein System logisch verknüpfter Aussagen zu einer Hypothese zu formulieren, die ausreichend präzise ist, um einer Überprüfung gegen die Realität standzuhalten.[46] Die Hypothesen stellen Zusammenhänge zwischen Systemkomponenten her und versuchen die Ursachen in Systemen zu erklären.[47] Grundlage für die Hypothesen sind Annahmen über die Gesetzmäßigkeiten der Ausgangslage.[48] Kausalmodelle gewinnen verstärkt an Bedeutung. Einerseits wegen der Möglichkeiten durch die neuen Entwicklungen in der multivariaten Statistik[49] und andererseits wegen der steigenden Komplexität durch die fortschreitende Digitalisierung.[50]

Prognosemodelle

Prognosemodelle sind eine Sonderform von Erklärungsmodellen, die Daten vorhersagen und Auswirkungen von Handlungsalternativen abschätzen.[51] Die Basis der Vorhersagen sind Daten und Ereignisse der Vergangenheit.[52]

Entscheidungsmodelle

Ein Entscheidungsmodell bewertet anhand einer Zielgröße Handlungsalternativen und wählt die Vorteilhafteste.[53] Der Unterschied zu Erklärungsmodellen, deren Ergebnisse als Entscheidungsgrundlage herangezogen werden können, ist, dass Entscheidungsmodelle so konstruiert sind, dass das Ergebnis in Form der Entscheidung sich direkt aus dem Modell ergibt.[54]

Simulationsmodelle

Simulationsmodelle sind eine Sonderform der Prognosemodelle, jedoch keine Erklärungsmodelle im engeren Sinn. Die Systeme, die das Modell abbildet, sind komplexer, mit stochastischen Komponenten und daher analytisch nicht einfach beschreibbar. Unterschiedliche Handlungsalternativen werden durchgespielt und auf ihre Zieleinwirkung hin untersucht. Es stellt eine Ergänzung zum Entscheidungsmodell für den Fall dar, dass die Lösung des Entscheidungsmodells zu aufwendig ist. Es wird am Ende jene Handlungsalternative gewählt, die sich auf die Ziele am günstigsten auswirkt.[55]

---

[45] Vgl. Adam, D. (1993), S. 71.
[46] Vgl. Homburg, C. (1998), S. 35.
[47] Vgl. Adam, D. (1993), S. 71.
[48] Vgl. Jockisch, M.; Rosendahl, J. (2010), S. 32.
[49] Vgl. Homburg, C. (1998), S. 35.
[50] Vgl. Bennett, N.; Lemoine, G. J. (2014), S. 313 ff.; Vgl. Budde, L.; Friedli, T. (2017), S. 37.; Scheller, T. (2017), S. 20 ff.
[51] Vgl. Jockisch, M.; Rosendahl, J. (2010), S. 32.
[52] Vgl. Homburg, C. (1998), S. 35.
[53] Vgl. Adam, D. (1993), S. 71.
[54] Vgl. Jockisch, M.; Rosendahl, J. (2010), S. 32.
[55] Vgl. Domschke, W.; Scholl, A. (2005), S. 31.

2.2.2 Alternative Modellklassifizierungen

Vor dem Hintergrund der breiteren Einsetzbarkeit des entwickelten Modells werden weitere Modellklassifikationen vorgestellt, die für eine Einteilung relevant sind.

Referenzmodelle

Das Software Engineering Institute (SEI) der Carnegie Mellon University definiert ein Referenzmodell als „Ein Modell, das als Maßstab für die Messung eines Attributs verwendet wird.“[56] Es ist eine spezielle Form von Informationsmodell[57] und daher in der Literatur auch unter dem Begriff Referenz-Informationsmodell zu finden. Es hat die Allgemeingültigkeit und den Empfehlungscharakter als definierendes Merkmal. Durch den Empfehlungscharakter bildet es einen Bezugspunkt in der Entwicklungsphase indem es eine Klasse von Anwendungsfällen umfasst. Die Referenz stellt den „best practice“ oder „good practice“ dar.[58]

Phasen- und Prozessmodelle

Modelle dieser Klassifikation benennen die begrifflichen, organisatorischen, technischen und zeitlichen Kriterien, in die eine Entwicklungsaufgabe gegliedert werden kann. Sie geben an was für eine Zielerreichung getan werden muss. Typische Phasen der Systementwicklung sind die Analyse, der Entwurf, die Realisierung und die Implementierung[59] oder die Phasen des Systems Engineerings.[60]

Vorgehensmodelle

Während Phasen- und Prozessmodelle die Frage des „Was?“ beantworten und die Teile eines Systems beschreiben, erweitern die Vorgehensmodelle die Sicht um das „Wie?“ und beschreiben das Verhalten der Komponenten. Sie beschreiben die Abfolge der Aktivitäten, die für eine Aufgabe notwendig sind und sie geben an wie die Aktivitäten oder Methoden anzuwenden sind.[61]

Kompetenzmodelle

AHLEMANN ET AL. definieren Kompetenzmodelle wie folgt:

„Ein Kompetenzmodell dient der Beurteilung, inwieweit ein Kompetenzobjekt die für eine Klasse von Kompetenzobjekten allgemeingültig definierten qualitativen Anforderungen erfüllt. Hierzu wendet ein Assessor Informationserhebungs- und Analysemethoden unter Einbeziehung von Informationslieferanten an. Das Ergebnis wird Modellempfängern für ihre Zwecke zur Verfügung gestellt.“[62]

---

[56] CMMI Product Team (2010), S. 411. (aus dem Englischen übersetzt)
[57] „Ein Informationsmodell (genauer: Informationssystemmodell) ist ein spezielles expliziertes Modell, dessen Gegenstand ein Informationssystem ist“ Vom Brocke, J.; Grob, H. L. (2015), S. 26.
[58] Vgl. Schütte, R. (1998), S. 69.; Vom Brocke, J.; Grob, H. L. (2015), S. 31 f.
[59] Vgl. Goeken, M. (2006), S. 51.; Vom Brocke, J.; Grob, H. L. (2015), S. 67 f.
[60] Siehe: Haberfellner, R. et al. (2012), S. 157.
[61] Vgl. Goeken, M. (2006), S. 52.; Vom Brocke, J.; Grob, H. L. (2015), S. 68.
[62] Ahlemann, F. et al. (2005), S. 14.

Ein Kompetenzobjekt hat Anforderungen und für eine Klasse von Kompetenzobjekten existiert ein Konsens über die Anforderungen. Mit einem Kompetenzobjekt wird ein Ausschnitt aus der Realität abgebildet. Der Assessor ist einer der Nutzer eines Kompetenzmodells. Er wendet es an und bewertet die Kompetenzobjekte mit Methoden der Informationsbeschaffung und –analyse.[63]

2.3 Reifegradmodelle

Zur Einordnung von Reifegradmodellen in eine der erwähnten Modellkategorien werden die Begriffe Modell und Reifegrad gesondert betrachtet. Der Modellbegriff wurde bereits definiert und Reifegradmodelle müssen zumindest die drei Merkmale – Abbildungsmerkmal, Verkürzungsmerkmal, pragmatischen Merkmal – erfüllen.[64] Ein Reifegrad ist das Ausmaß der Prozessverbesserung in einer Menge an vordefinierten Prozessschritten oder Prozessbereichen. Alle Punkte, die einen Reifegrad ausmachen, müssen erfüllt sein, damit dieser als erreicht gilt.[65] Die Reifegrade bauen aufeinander auf. Eine weit verbreitete generische Semantik ist:[66]

1. Reifegrad: Chaotische Prozesse
2. Reifegrad: Definierte Prozesse
3. Reifegrad: Standardisierte Prozesse
4. Reifegrad: Gemessene Prozesse
5. Reifegrad: Kontinuierliche Verbesserung

Reifegradmodelle könnten den Beschreibungsmodellen zugeordnet werden. Sie beschreiben den aktuellen Zustand von Systemelementen, jedoch nicht die Beziehung untereinander. Mit den Reifegraden stellen sie einen Sollzustand dar, was für eine Einteilung als Referenzmodell spricht.

Die Definition der Kompetenzmodelle trifft die Eigenschaften und die Anwendungsmethodik von Reifegradmodellen gut. AHLEMANN ET AL. haben die Definition formuliert, da es deren Meinung nach keine konkreten Definitionen von Reifegradmodellen gibt.[67] Für sie sind Reifegradmodelle eine Sonderform der Kompetenzmodelle.[68] Sie bauen ihre Definition jedoch auf der Definition von Reifegradmodellen laut SEI auf, die ein Reifegradmodell als Referenzmodel definieren.

Reifegradmodelle sind laut SEI folgend definiert:

„A model that contains the essential elements of effective processes for one or more areas of interest and describes an evolutionary improvement path from ad hoc, immature processes to disciplined, mature processes with improved quality and effectiveness."[69]

In dieser Arbeit wird das Reifegradmodell den Referenzmodellen zugeordnet. Es wird so aufgebaut, dass der datenanalytische Prozess anhand von greifbaren Bereichen – in

[63] Vgl. Ahlemann, F. et al. (2005), S. 13 f.
[64] Vgl. Jochem, R. (2010), S. 115.
[65] Vgl. CMMI Product Team (2010), S. 401. (übersetzt aus dem Englischen)
[66] Vgl. Ahlemann, F. et al. (2005), S. 19.
[67] Vgl. Ahlemann, F. et al. (2005), S. 12.
[68] Vgl. Ahlemann, F. et al. (2005), S. 14.
[69] CMMI Product Team (2010), S. 436.

diesem Fall Kategorien – bewertet wird. Der Konnex zwischen den Bewertungskategorien (Reifegradkategorien) und der Beurteilungsskala (Reifegrade) solle es erlauben, die Bedingungen für den Prozess kontinuierlich zu verbessern, um diesen durch Verbesserungen in den Bewertungskategorien effizienter durchführen zu können und die Qualität seines Outputs zu erhöhen.

2.3.1 Arten von Reifegradmodellen

Seit der Veröffentlichung des ersten Reifegradmodells, dem Capability Maturity Model (CMM) des SEI der Carnegie Mellon University wurden in ähnlicher Form eine große Zahl von Reifegradmodellen veröffentlicht. Folgend werden ausgewählte Reifegradmodelle genauer vorgestellt. Wobei in jene unterschieden wird, die weitgehend als Standard gelten und neuere, die eng abgegrenzte Themenbereiche betreffen und aufgrund der kurzen Zeit seit deren Veröffentlichung noch nicht über die Bekanntheit verfügen, um als Standard zu gelten.

Quality Management Maturity Grid

Viele Reifegradmodelle, die in den letzten Jahrzehnten veröffentlicht wurden, basieren in deren Kern auf dem Quality Management Maturity Grid von CROSBY[70], welches dieser 1979 erstmals vorstellte. Die Reifematrix zeigte die Reifephase von Unternehmen im Qualitätsmanagement. Dem Management soll verdeutlicht werden, welche Maßnahmen im Qualitätsmanagement ursprünglich nötig sind, wo sich das Unternehmen im Vergleich zum Ideal befindet und welche Tätigkeiten folglich initiiert werden müssen, um dem Ideal näher zu kommen.[71] Tabelle 3 zeigt das Quality Management Maturity Grid. Sechs Handlungsfelder des Qualitätsmanagements werden in fünf Bewertungsstufen beurteilt.[72]

Tabelle 3: Quality Management Maturity Grid[73]

Kategorien zur Messung der Qualität	Stadium I: Unsicherheit	Stadium II: Erwachen	Stadium III: Erkenntnis	Stadium IV: Verständnis	Stadium V: Sicherheit
Qualitäts-verständnis	Wird nicht als Management-instrument gesehen	Einsicht, jedoch keine Bereitschaft Geld oder Zeit zu investieren	Einsicht, dass man Unterstütz-ung geben muss	Beteiligung und Akzeptanz der Verpflichtung für Engagement	QM als unverzichtbarer Bestandteil des Unternehmens-systems verstanden
Status der Qualitäts-organisation	Verantwortung in der Fertigung oder Entwicklung. Fokus liegt auf Bewerten und Sortieren	Stärkerer Qualitätsleiter ernannt. Keine weitere Änderung zu Stadium I	Qualitäts-abteilung; integrierte Bewertungs-methoden; Qualitätsleiter in die Führung einbezogen	Qualitäts-manager auf Führungsebene; Statusberichte und Vorbeuge-maßnahmen	Qualitäts-manager in der Geschäfts-leitung; Vorbeugen liegt im Fokus; Qualitätsdenken

[70] siehe Originalveröffentlichung: Crosby, P. B. (1980)
[71] Vgl. Crosby, P. B. (2000), S. 49.
[72] Vgl. Jochem, R. (2010), S. 116.
[73] Quelle: Crosby, P. B. (2000), S. 50 f. leicht modifiziert

Fortsetzung zu Tabelle 3: Quality Management Maturity Grid[74]

Problem-behandlung	Probleme erst beim Auftreten bekämpft; keine endgültigen Lösungen	Vorrangige Probleme von Teams behandelt; keine langfristigen Lösungen gesucht	Kommunikation der Maßnahmen; Systematische Problemlösung	Probleme im Entwicklungs-stadium identifiziert; Offenheit gegenüber Vorschlägen	Problemen wird vorgebeugt
Qualitätskosten in % des Umsatzes	Laut Bericht: Unbekannt Realität: 20%	Laut Bericht: 3% Realität: 18%	Laut Bericht: 8% Realität: 12%	Laut Bericht: 6,5% Realität: 8%	Laut Bericht: 2,5% Realität: 2,5%
Maßnahmen zur Qualitäts-verbesserung	Keine Koordinierten, und kein Verständnis dafür	Oberflächlich und kurzfristig	Implementierung eines Prozesses und Verständnis dafür	Fortsetzung des Prozesses; Qualität wird in die Hauptprozesse aufgenommen	Qualitäts-verbesserung ist eine normale und kontinuierliche Aktivität
Zusammen-fassung der Qualitäts-einstellung des Unternehmens	„Wir wissen nicht, warum wir Probleme mit der Qualität haben.“	„Sind ständige Probleme mit der Qualität absolut unvermeidlich?“	„Durch Engagement seitens des Managements und der Qualitäts-verbesserung gelingt es uns, unsere Probleme zu identifizieren und zu lösen.“	„Fehlerver-meidung ist ein fester Bestandteil unserer Tätigkeit.“	„Wir wissen, warum wir keine Probleme mit der Qualität haben.“

Capability Maturity Model (CMM)

CMM ist das Reifegradmodell, auf welchem viele der später entwickelten basieren. Das entscheidende Wiedererkennungsmerkmal sind die fünf aufeinander aufbauenden Reifgrade die generisch definiert wurden und auf andere Bereiche übertragbar sind.[75]

CMM wurde speziell für IT- und Softwareentwicklungsprozesse mit dem Ziel entwickelt, einen Reifegrad zu bestimmen, der eine Aussage über die Durchführbarkeit von IT-Prozesse in den Bereichen Definiertheit, Dokumentation, Planung, Steuerung und Kontrolle treffen kann. Das Profil des Reifegrades ist eine Messlatte für einen Stärken-Schwächen-Vergleich. Dieser Vergleich ist in weiterer Folge ein Benchmark für die Ableitung von Verbesserungsmaßnahmen zur Weiterentwicklung der bewerteten IT-Prozesse. Es wird in fünf Reifegradstufen unterschieden, zu welchen eine Prozesscharakteristik beschrieben wurde.[76] Tabelle 4 gibt einen Überblick der Reifegrade und beschreibt die dazugehörige Prozesscharakteristik.

[74] Quelle: Crosby, P. B. (2000), S. 50 f. leicht modifiziert
[75] Vgl. Jochem, R. (2010), S. 116.
[76] Vgl. Harrach, H. (2010), S. 35.

Tabelle 4: Reifegrade nach CMM[77]

	Reifegrad	Prozesscharakteristik
1	Initial (initial)	Prozesse ad-hoc in chaotischer Umgebung. Erfolg durch Heldentaten Einzelner; häufig nicht wiederholbar.
2	Wiederholbar (repeatable)	Einfaches Projektmanagement für Zeit- und Kostenüberwachung. Prozessdisziplin für etablierte Verfahren. Erfolge wiederholbar.
3	Definiert (defined)	Firmeninterne Prozesse sind verstanden und werden in Standards und Prozeduren umgesetzt. Sind auf den jeweiligen Kontext zugeschnitten.
4	Verwaltet (managed)	Maße für die Prozessvorhersage sind definiert. Prozess ist übertragbar.
5	Optimierend (optimizing)	Prozessleistung wird stetig verbessert. Lernen wird umgesetzt. Anpassungen werden weitgehend automatisiert.

Je höher die Reife, desto besser ist die Prozessfähigkeit eines Unternehmens. Dies wirkt sich auf die Datenqualität aus, da der Prozess zur Sicherstellung der Datenqualität durch Regeln, Standards und organisatorische Strukturen institutionalisiert wird. CMM fordert, dass alle Ziele eines Reifegrades und jene der Reifegrade davor erfüllt werden, um einen Reifegrad zu erreichen.[78]

Capability Maturity Model Integration (CMMI)

Das CMMI ist eine Weiterentwicklung des CMM, welches ebenfalls vom SEI entwickelt wurde. Das Modell erweitert den reinen Softwarefokus der Vorgängerversion.[79] Dadurch ist es flexibler einsetzbar und wurde in einer Vielzahl neuer Modelle abgewandelt.

Die fünf Reifegrade sind:[80]

1. Reifegrad: Initial (Ad hoc Prozesse)
2. Reifegrad: Gemanagt (geplante Prozesse)
3. Reifegrad: Definiert (verstandene und standardisierte Prozesse)
4. Reifegrad: Quantitativ gemanagt (Ziele und Qualität festgelegt)
5. Reifegrad: Optimierend (ständig verbesserte Prozesse)

Reifegrad Initial:

Prozesse in diesem Reifegrad werden chaotisch ausgeführt. Mitarbeiter sind durch ihren persönlichen Einsatz für das Funktionieren der Organisation verantwortlich, da sie nicht die Grundlage für ein stabiles Funktionieren der Prozesse liefert. Dieses offensichtliche Chaos bedeutet jedoch nicht, dass das Unternehmen nicht funktioniert und keine Leistung liefert, sondern dass es vermehrt zu Budget- und Zeitüberschreitungen kommt. Unternehmen in diesem Reifegrad geben ihre Prozesse in Stresssituationen leicht auf und können Erfolge nicht wiederholen.[81]

[77] Quelle: in Anlehnung an Paulk, M. C. et al. (1995), S. 8 f.
[78] Vgl. Harrach, H. (2010), S. 37.
[79] Vgl. Jochem, R. (2010), S. 126 f.
[80] Vgl. Chrissis, M. B. et al. (2003), S. 101 ff.
[81] Vgl. CMMI Product Team (2010), S. 29.

Reifegrad Gemanagt:

Reifegrad zwei zeichnet sich durch ein umgesetztes Projektmanagement aus. Es bildet die Grundlage für den geordneten Ablauf von Prozessen durch die Erstellung von Projektplänen, die Überwachung und Steuerung der Projekte. Die Prozessleistung kann gemessen und analysiert werden. Projekte, Prozesse und Dienstleistungen sind unter Kontrolle oder gemanagt. Relevante Stakeholder werden identifiziert und eingebunden. Um die Prozesse durchzuführen, werden ausreichend Ressourcen zu Verfügung gestellt. Dadurch wird eine Grundlage geschaffen, die gelebten und etablierten Praktiken unter Krisenbedingungen weiter auszuführen.[82]

Reifegrad Definiert:

Der dritte Reifegrad zeichnet sich dadurch aus, dass definierte Prozesse verwendet werden. Es wird überprüft ob Erzeugnisse oder Dienstleistungen den Anforderungen entsprechen. Die angewandten Prozesse sind beschrieben, verstanden und standardisiert. Standardprozesse sind definiert und werden laufend verbessert. Sie dienen zur Förderung der Vereinheitlichung in der Organisation. Im Unterschied zu Reifegrad zwei werden die Prozesse für Projekte nicht jeweils neu definiert, sondern ergeben sich aus den Standardprozessen des Unternehmens.[83]

Reifegrad Quantitativ gemanagt:

Im vierten Reifegrad werden quantitative Qualitäts- und Leistungsziele definiert. Die Ziele werden von den Anforderungen der Prozessstakeholder abgeleitet. Die Ziele, Messungen und Analysen werden auf wichtige Subprozesse angewandt. Dafür sind die Wirkungsbeziehungen der Subprozesse untereinander zu bestimmen, sowie deren Einfluss auf die Ziele des Hauptprozesses. Reifegrad vier unterscheidet sich von Reifegrad drei durch die verbesserte Vorhersagbarkeit der Prozessleistung.[84]

Reifegrad Optimierend:

Als höchster Reifegrad hat „Optimierend“ eine laufende Verbesserung der Prozesse gemessen an den Geschäftszielen zum Ziel. Die Verbesserung wird durch eine schrittweise und innovative Verbesserung der Prozesse und der eingesetzten Technologien erreicht. Die Qualitäts- und Leistungsziele der Organisation sind definiert und bekannt und werden in regelmäßigen Abständen evaluiert. Die Verbesserungen werden gemessen und mit Zielwerten verglichen, um eine laufende Weiterentwicklung zu gewähren. Der wesentliche Unterschied zu Reifegrad vier ist der holistische Blick auf den Unternehmensprozess. Reifegrad vier konzentriert sich auf die Subprozesse. In Reifegrad fünf werden unterschiedliche Daten analysiert und Schwächen in der Leistung identifiziert. Diese Schwachstellen sind der Ausgangspunkt für eine weitere messbare Verbesserung im Leistungsprozess.[85]

[82] Vgl. CMMI Product Team (2010), S. 29 f.
[83] Vgl. CMMI Product Team (2010), S. 30.
[84] Vgl. CMMI Product Team (2010), S. 30 f.
[85] Vgl. CMMI Product Team (2010), S. 31.

Zusätzlich zu den Reifegraden wurden vier (0 bis 3) Fähigkeitsgrade – unvollständig (0), durchgeführt (1), gemanagt (2) und definiert (3) – festgelegt, die die Art der Prozessdurchführung beschreiben.

Ein unvollständiger Prozess wird nicht oder nur teilweise ausgeführt. Spezifische Ziele werden nicht vollständig erfüllt und generische Ziele sind nicht definiert. Ein Prozess des Fähigkeitsgrades „durchgeführt“ erfüllt die erwartete Leistung und liefert den definierten Output. Spezifische Ziele werden auf dieser Stufe erfüllt. Der Prozess ist nicht institutionalisiert und kann daher über die Zeit auf Fähigkeitsgrad „0“ zurückfallen. Als gemanagt gilt ein Prozess mit dem Fähigkeitsgrad „durchgeführt“, wenn er geplant ist und entsprechend der Vorschriften durchgeführt wird. Dem Prozess steht qualifiziertes Personal und ausreichend Ressourcen zur Verfügung und er produziert ein überprüfbares Ergebnis. Der höchste Fähigkeitsgrad „definiert“ wird erreicht, wenn ein gemanagter Prozess speziell für eine Organisation entsprechender standardisierter Vorgaben abgeleitet wurde. Die Prozessbeschreibung wird am aktuellen Stand gehalten und der Prozess trägt zu den Prozessassets bei. Prozesse der Fähigkeitsstufe 3 sind präziser beschrieben als jene des zweiten Grades.[86]

ISO/IEC 15504 – Software Process Improvement and Capability Determination

SPICE wurde als europäische Antwort auf CMMI entwickelt. Das Modell dient zur Bewertung von Unternehmensprozessen für die Softwareentwicklung. Aus der Erstveröffentlichung im Jahre 1998 entwickelte sich die internationale Norm ISO/IEC 15504 im Jahr 2006. Die Bewertung der Prozesse ist weniger komplex als bei CMMI. Auf konkrete Handlungsanweisungen wurde zu Gunsten der Möglichkeit einer unternehmensindividuellen Anpassung verzichtet. Grundlage dafür sind Erfahrungswerte und zu erwartende Ergebnisse.[87]

SPICE hat für die Softwareentwicklung die Schwerpunkte *Prozesse bewerten und verbessern* und die *Bewertung des Prozessfähigkeitsgrads* bestimmt. Die implizierte Annahme der Entwickler ist, dass gute Prozesse die Wahrscheinlichkeit für gute Produkte und Dienstleistungen erhöht. Der Zyklus der Prozessverbesserung mit SPICE ist in Abbildung 3 graphisch dargestellt.

So genannte Assessoren führen im Rahmen eines Assessment die Bewertung der Prozesse durch. Dabei werden die Stärken und Schwächen mithilfe von Fragebögen und Interviews erhoben. Die Bewertung erfolgt auf einer sechs-stufigen Reifegradskala, die in SPICE Fähigkeitsdimensionen genannt werden.[88] Jeder Fähigkeitsstufe werden Prozessattribute zugeordnet die für die Erfüllung des Reifegrades erreicht werden müssen.[89] Tabelle 5 fasst die Reifegradstufen mit einer Beschreibung der Fähigkeitsdimension und den zugeordneten Prozessattributen in einer kompakten Übersicht zusammen.

[86] Vgl. CMMI Product Team (2010), S. 26 f.
[87] Vgl. Jochem, R. (2010), S. 127 f.
[88] Vgl. Siebert, H. G. (2010), S. 76.
[89] Vgl. Grande, M. (2011), S. 115.

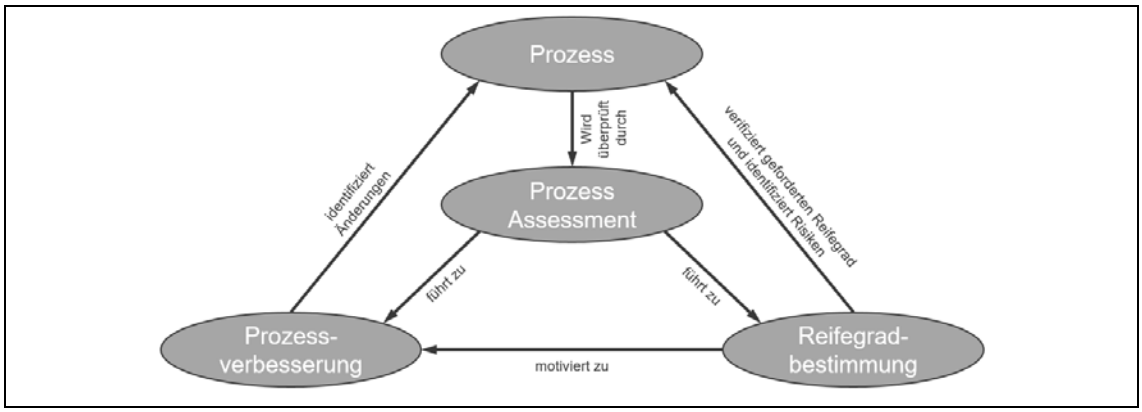

Abbildung 3: Prozessverbesserung mit SPICE[90]

Den Prozessattributen werden Managementaufgaben zugewiesen, die, wenn korrekt ausgeführt, den Prozess systematisch bearbeiten und die Ergebnisse in einer vordefinierten Qualität liefern.[91]

Tabelle 5: SPICE Reifegradstufen[92]

Stufe	Fähigkeitsdimension	Beschreibung	Prozessattribute
0	Unvollständiger Prozess (incomplete)	Der Prozess ist nicht implementiert oder verfehlt seinen Zweck.	
1	Durchgeführter Prozess (performed)	Der Zweck des Prozesses wird erfüllt. Der Prozess ist eingeführt und erfüllt die Prozessziele.	• PA 1.1 Prozess-durchführung
2	Gesteuerter Prozess (managed)	Die Ausführung des Prozesses wird geplant und gesteuert. Arbeitsprodukte sind festgelegt, werden kontrolliert und gepflegt.	• PA 2.1 Prozess-management • PA 2.2 Arbeits-erzeugnisse
3	Etablierter Prozess (established)	Ein definierter Prozess wird benutzt, der auf einem Standardprozess basiert.	• PA 3.1 Prozess-definition • PA 3.2 Prozess-anwendung
4	Vorhersagbarer Prozess (predictable)	Der Prozess wird konsistent ausgeführt, innerhalb definierter Grenzen.	• PA 4.1 Prozess-messung • PA 4.2 Prozess-steuerung
5	Optimierender Prozess (optimizing)	Der Prozess wird kontinuierlich verbessert, um relevante aktuelle und projektierte Geschäftsziele zu erreichen.	• PA 5.1 Prozess-innovation • PA 5.2 Prozess-optimierung

[90] Quelle: Siebert, H. G. (2010), S. 77.
[91] Vgl. Siebert, H. G. (2010), S. 80.
[92] Quelle: in Anlehnung an Siebert, H. G. (2010), S. 79.

2.3.2 Aufbau und Entwicklung von Reifegradmodellen

Reifegradmodelle bewerten direkt oder indirekt Prozesse. Zur Umsetzung wird ein Prozess in Schritte und Teil- und Obergebiete unterteilt. Die Bewertung erfolgt auf Ebene der Teilgebiete aus deren Gesamtbewertung sich die Bewertung der Obergebiete ergibt. Die Gesamtbewertung der Obergebiete ergibt den Reifegrad des Prozesses. Dieses grobe Vorgehen, lässt sich als Nukleus in allen Reifegradmodellen finden, die sich in vielen Fällen in ihrer Struktur und dem inhaltlichen Aufbau unterscheiden.

Die strukturelle Zusammensetzung von Elementen eines Reifegradmodells hat eine wesentliche Bedeutung in deren Ausgestaltung. Es lassen sich vier Ebenen unterscheiden, die *theoretische Ebene*, die *empirische Ebene*, die *Ebene der Bewertung* und die *Ebene der Reifebestimmung*. Abbildung 4 zeigt den hierarchischen Aufbau der Ebenen.

Auf der theoretischen Ebene werden die zu messenden Prozessschritte als Konstrukte definiert, welche in Subkonstrukte unterteilt werden, wobei Subkonstrukte Dimensionen haben können.[93] Für die Messung der Konstrukte muss eine Skala entwickelt werden, da diese nicht direkt messbar sind.[94] Inhalt der empirischen Ebene ist die Messung der Subkonstrukte über Indikatoren und Einzelindikatoren. Die Einzelindikatoren operationalisieren die Inhalte der Subkonstrukte, ermöglichen dadurch die Umsetzung und Durchführung des Prozesses, der durch das Konstrukt dargestellt ist. Durch die Messung oder Bewertung der Einzelindikatoren wird die Messung eines abstrakten Konstruktes möglich.[95]

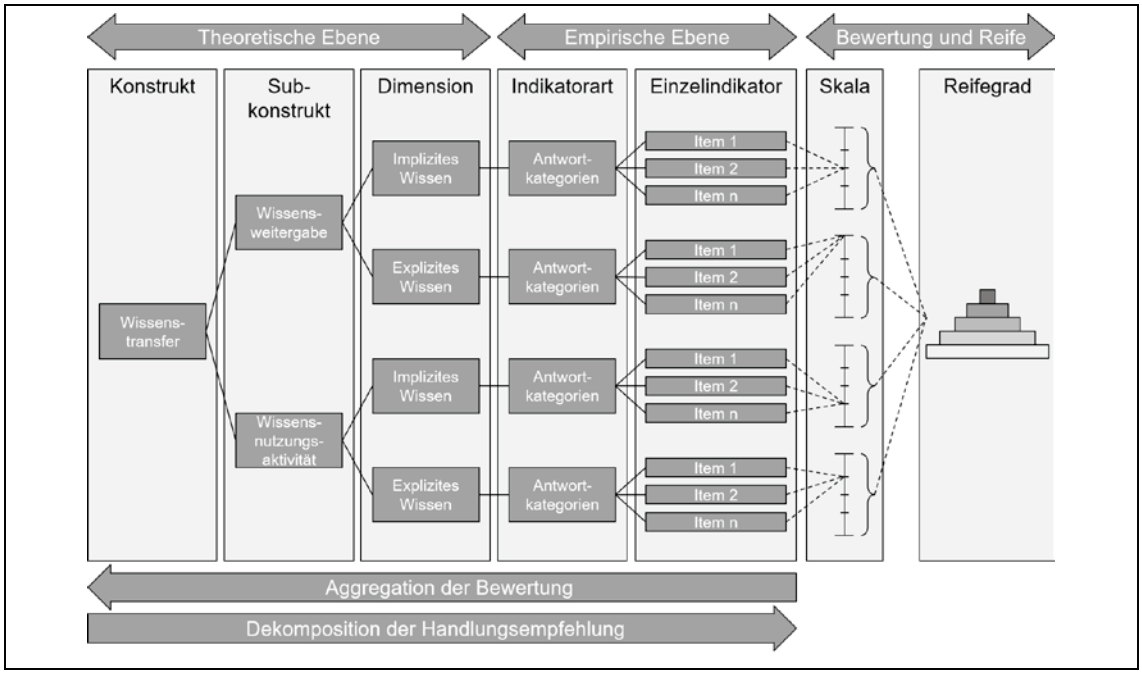

Abbildung 4: Aufbau und Struktur eines Reifegradmodells[96]

[93] Vgl. Jochem, R. (2010), S. 118 f.
[94] Vgl. Diekmann, A. (2001) zitiert nach Werner, M. (2004), S. 82.
[95] Vgl. Werner, M. (2004), S. 112.; Jochem, R. (2010), S. 118 f.
[96] Quelle: in Anlehnung an Jochem, R. (2010), S. 118.

Auf der empirischen Ebene wird der Erfüllungsgrad der Subkonstrukte und folglich des Konstruktes bezogen auf die vordefinierte Reifegradskala bestimmt. Die Erhebung erfolgt mit Fragebögen, Interviews, der Analyse relevanter Dokumente[97] oder durch die Analyse von Daten.

Der Inhalt der Erhebungsmethoden ergibt sich aus den Anforderungen der theoretischen Ebene. Dadurch wird der Konnex zwischen der theoretischen und empirischen Ebene hergestellt.[98] Die Skala auf der Bewertungsebene orientiert sich an den Ausprägungen der Reifegradstufen.

Die Antworten in den Einzelindikatoren, bzw. deren Vorhandensein, bestimmt die Einordnung auf der Skala, welche in Summe die Einordnung des Gesamtindikators bestimmt. Folglich ist die Bewertung der Dimensionen, der Subkonstrukte und final der Konstrukte möglich.[99] Durch diesen hierarchisch aggregierten Bewertungsaufbau kann eine Vielzahl von Einzelindikatoren auf wenige Reifegradkategorien zusammenzufassen werden. Es können des Weiteren spezifische Handlungsempfehlungen für die Reifegradverbesserung ausgesprochen werden,[100] indem eine Dekomposition entlang der Hierarchie vorgenommen wird und am Ende, wenn nötig, die Einzelindikatoren mit dem meisten Potenzial aufzuzeigen.

CMMI gibt 16 Kernprozessbereiche[101] vor, die für eine Prozessverbesserung in jeder Umgebung grundlegend sind.[102] Abhängig vom Einsatzgebiet für das das Reifegradmodell mit dem CMMI Rahmenwerk erstellt wird werden die Kernprozessbereiche um weitere Prozessbereiche erweitert.[103]

Prozessbereich haben spezifische und generische Ziele. Spezifische Ziele sind dem Prozessbereich eigen und müssen erfüllt werden, um dem Prozessbereich zu genügen. Generische Ziele können in mehreren Prozessbereichen vorkommen. Die Ziele, bzw. die Voraussetzungen für die Zielerfüllung müssen in der Erhebung der Fähigkeit oder Reife abgefragt werden. Die entspricht in Abbildung 4 den Indikatoren der empirischen Ebene.

Um die spezifischen Ziele erreichen zu können, werden spezifische Praktiken beschrieben. Für die Praktiken werden Outputs definiert, anhand derer die korrekte Durchführung der Aktivitäten, die mit den Praktiken assoziiert werden, überprüft werden kann. Um die Aktivitäten durchzuführen sind u. U. Voraussetzungen oder Deliverables von Prozesslieferanten nötig. Die können ebenfalls definiert und überprüft werden. Die Feststellung der Reife über die Praktiken, Outputs und Deliverables ermöglicht die in Abbildung 4 dargestellte Dekomposition von Handlungsempfehlungen, um Fähigkeitsgrade und Reifegradstufen zu erreichen. Für die generischen Ziele gibt es analog zum beschriebenen Vorgehen generische Praktiken, die durchgeführt werden, um die Ziele zu erreichen.[104]

---

[97] Vgl. Jochem, R. (2010), S. 119.
[98] Vgl. Jochem, R. (2010), S. 119.
[99] Vgl. Werner, M. (2004), S. 114.
[100] Vgl. Werner, M. (2004), S. 215.
[101] Siehe Linders, B. (2019)
[102] Vgl. CMMI Product Team (2010), S. 11.
[103] Vgl. CMMI Product Team (2010), S. 3 f.
[104] Vgl. CMMI Product Team (2010), S. 14 ff.

Im Namen von CMMI ist eine grundlegende Unterscheidung zu finden, die Reifegradmodelle nach diesem Rahmenwerk in deren Aufbau und Bewertung treffen. Zum einen werden die Fähigkeit eines Prozessbereich[105] (Capability) und zum anderen die Reife (Maturity) des gesamten Prozesses innerhalb der Organisation unterschieden.

Verbesserungen können im Rahmen der Fähigkeit oder der Reife erfolgen. Im ersten Fall spricht man von der kontinuierlichen Darstellung, im zweiten Fall von der stufenweisen Darstellung der Verbesserung. Um Fähigkeitsgrade oder Reifegradstufen zu erreichen, müssen alle Ziele der Prozessbereiche erfüllt werden.[106]

Die stufenweise Darstellung stellt den Gesamtzustand der Prozesse eines Unternehmens im Vergleich zum Gesamtmodell dar. Die kontinuierliche Darstellung verwendet die Fähigkeitsgrade, um den Zustand der Prozesse im Unternehmen im Verhältnis zu einem Prozessbereich zu beschreiben.[107]

Tabelle 6 zeigt eine Gegenüberstellung der Fähigkeits- und Reifegrade. Ziel der kontinuierlichen Darstellung ist die Wahl und die Verbesserung eines Prozessbereiches. Daher ist es wichtig ob der Prozess durchgeführt wird oder unvollständig ist. Die stufenweise Darstellung will mit der Auswahl mehrerer Prozessbereiche die Reife verbessern. Die Unterscheidung ob ein Prozess durchgeführt wird oder unvollständig ist, ist dafür nicht wichtig, daher startet die Reifegradbewertung bei Grad eins – initial.[108]

Tabelle 6: Vergleich Fähigkeits- und Reifegrade[109]

Grad	Kontinuierliche Darstellung Fähigkeitsgrade	Stufenweise Darstellung Reifegrade
Grad 0	Unvollständig	
Grad 1	Durchgeführt	Initial
Grad 2	Gemanagt	Gemanagt
Grad 3	Definiert	Definiert
Grad 4		Quantitativ gemanagt
Grad 5		Optimierend

Der Zusammenhang zwischen den beiden Darstellungsformen ergibt sich durch die Durchführung der Prozessverbesserung. Mit Fähigkeitsgraden ist es Organisationen möglich spezifische Prozessbereiche zu wählen die verbessert werden sollen. Deren IST-Ausprägung kann festgestellt werden und die SOLL-Ausprägung wird mit einem Zielprofil definiert. So kann ein Prozessbereich auf Fähigkeitsstufe drei einwickelt werden, während ein anderer nur Fähigkeitsstufe eins hat. Das CMMI-Rahmenwerk gibt an welche Kernprozessbereiche in einer Organisation entwickelt sein müssen, um einen

[105] Ein Prozessbereich ist eine Sammlung ähnlicher Praktiken eines Bereichs. Werden diese umgesetzt führen sie zu einer Zielerfüllung die wichtig für die Verbesserung im Bereich ist. (Vgl. CMMI Product Team (2010), S. 13.)
[106] Vgl. CMMI Product Team (2010), S. 23 f.
[107] Vgl. CMMI Product Team (2010), S. 24.
[108] Vgl. CMMI Product Team (2010), S. 25 f.
[109] Quelle: CMMI Product Team (2010), S. 25. (aus dem Englischen übersetzt)

bestimmten Reifegrad zu erreichen. Für die entwickelten Reifegradmodelle werden ggf. weitere Prozessbereiche definiert und Reifegradstufen zugeordnet. Die Prozessbereiche werden somit nach Reifegraden gruppiert. Die Zuordnung ist dem CMMI for acquisition, Version 1.3 zu entnehmen[110].

Die Zuordnung von Prozessattributen zu Fähigkeitsstufen im SPICE-Modell in Tabelle 5 kann damit verglichen werden. Ein Fähigkeitsstufenprofil kann in ein Reifegradprofil umgewandelt werden. Um Reifegrad zwei zu erreichen, müssen alle Prozessbereiche, die dem Reifegrad zugeordnet sind, mindestens den Fähigkeitsgrad zwei erreichen. Für Reifegrad drei müssen die Prozessbereiche von Reifegrad zwei und Reifegrad drei wie Fähigkeitsstufe drei haben. Analog setzt sich die Logik bis Reifegradstufe fünf fort. Es zeigt sich somit, dass es kein Zufall in der Definition ist, dass die Fähigkeitsgrade zwei und drei gleich benannt sind wie die Reifegrade zwei und drei. Um einen Reifegrad der Stufe „Definiert“ zu erreichen, genügt es nicht, nur die nötigen Prozessbereiche in der Organisation zu implementieren, für alle Reifegrade ab drei müssen die Prozesse einen Fähigkeitsgrad von drei erfüllen. Dieser ist wie bereits beschrieben nicht nur präzise definiert, sondern dem Umfeld und dem Einsatzzweck nach einer allgemeingültigen Vorgehensweise speziell angepasst worden.[111] Abbildung 5 zeigt den Zusammenhang der Bewertung einzelner Prozessbereiche auf Basis der Fähigkeitsgrade und der Zusammenführung zur Festlegung des Reifegrades eines Prozesses. Als Grundstruktur wurde jene von Abbildung 4 gewählt.

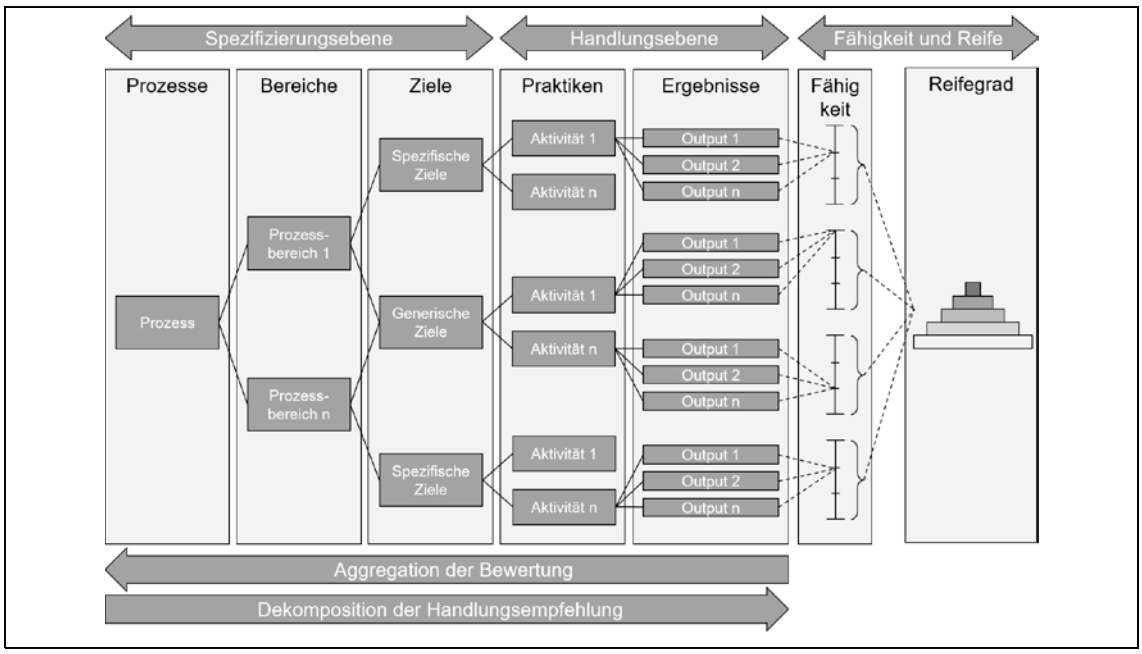

Abbildung 5: Bewertungsumgebung CMMI[112]

Es zeigt sich, dass ein gut strukturiertes Reifegradmodell grob in drei Bereiche unterteilbar ist. Der erste Bereich legt das theoretische Fundament, auf Basis dessen

[110] Vgl. CMMI Product Team (2010), S. 35 f.
[111] Vgl. CMMI Product Team (2010), S. 33 ff.
[112] Quelle: Eigene Darstellung, in Anlehnung an CMMI Product Team (2010), S. 12.; CMMI Product Team (2010), S. 24.; Jochem, R. (2010), S. 118.

eine Bewertung durchgeführt wird. Diese Basis dient zur Vergleichbarkeit und Standardisierung über die Grenzen eines Unternehmens oder einer Organisation hinweg. Der zweite Bereich bricht die theoretische Fundierung in die praktische Umsetzung herunter. Hier wird die Grundlage für die Bewertung gelegt anhand derer das Tun der Unternehmen bewertet wird. Der dritte Bereich ist die Bewertung selbst. Hier wird die Logik der Bewertungsskala festgelegt und die Zusammenführung zu einem Gesamtreifegrad definiert. Dieses Prozedere hat den Vorteil, dass das Vorgehen des Unternehmens auf eine vergleichbare und objektive Ebene aggregiert werden kann. Dadurch ist es weiter möglich spezifische Handlungsempfehlungen zu geben, um einen höheren Reifegrad zu erreichen. Dies erlaubt die definierte Liste an bewertbaren Ergebnissen, Outputs oder anderer Indikatoren und auch der Vergleich von Organisationen mit ähnlichen Reifegradprofilen im Bereich der theoretischen Fundierung.

2.3.3 Reifegradmodelle der Digitalisierung

Durch die Industrie 4.0 Initiative wurden in den letzten Jahren vermehrt Reifegradmodelle zur Bewertung der Facetten von Industrie 4.0 veröffentlicht. Folgend werden exemplarisch einige der Modelle vorgestellt.

Industrie 4.0 Maturity Index

Ziel ist die Feststellung des Reifegrades eines Unternehmens in der Umsetzung von Industrie 4.0. Basierend auf dem ermittelten Reifegrad sollen den Unternehmen Handlungsempfehlungen gegeben werden[113]. Die Reife wird in vier Gestaltungsfeldern und sechs Reifegradstufen bestimmt. Die Reifegradstufen ergeben sich aus den sechs Stufen des Industrie 4.0-Entwicklungspfades. Diese sind:[114]

1. Computerisierung
    - Auf dieser Ebene sind Informationstechnologien noch als Insellösungen implementiert.
    - Daten müssen z. B. händisch übertragen werden.
2. Konnektivität
    - Insellösungen werden abgelöst und Komponenten werden vernetzt.
    - Die vollständige Integration von IT[115] und OT[116] ist jedoch noch nicht vollzogen.
3. Sichtbarkeit
    - Sensoren erfassen entlang der internen Supply Chain alle relevanten Daten.
    - Ein digitales Modell des Unternehmens wird erzeugt, der auch digitaler Schatten genannt wird.
4. Transparenz
    - Das Unternehmen ist als digitaler Schatten abgebildet.

[113] Vgl. Schuh, G. et al. (2017), S. 13.
[114] Vgl. Schuh, G. et al. (2017), S. 15 ff.
[115] IT: Informationstechnologie
[116] OT: Operative Technologie. Hard- und Software die am Shopfloor an Anlagen und Geräten zum Einsatz kommt

- Mit den Daten werden Ursachenanalysen durchgeführt und Wirkungszusammenhänge interpretiert.

5. Prognosefähigkeit
   - Der digitale Schatten wird mit Simulationsmodellen in die Zukunft projiziert.
   - Durch die gewonnenen Szenarien können Störungen vorab erkannt und vermieden werden.
6. Adaptierbarkeit
   - Entscheidungen werden vom IT-System autonom getroffen.
   - Die Daten werden so eingesetzt, dass die positiven Effekte für das Unternehmen ohne menschliches Zutun maximiert werden.

Die vier Gestaltungsfelder sind Ressourcen, Informationssysteme, Organisationsstruktur und Kultur. Für die Gestaltungsfelder werden jeweils zwei handlungsleitende Prinzipien und dazugehörige Fähigkeiten identifiziert.[117] Abbildung 6 zeigt den Aufbau des Bewertungsschemas. Die Achsen der Viertelkreissegmente stellen die beiden handlungsleitenden Prinzipien je Gestaltungsfeld dar. Die sechs konzentrischen Ringe entsprechen den sechs Reifegraden. Für jedes Gestaltungsfeld erfolgt die Reifegradbewertung.

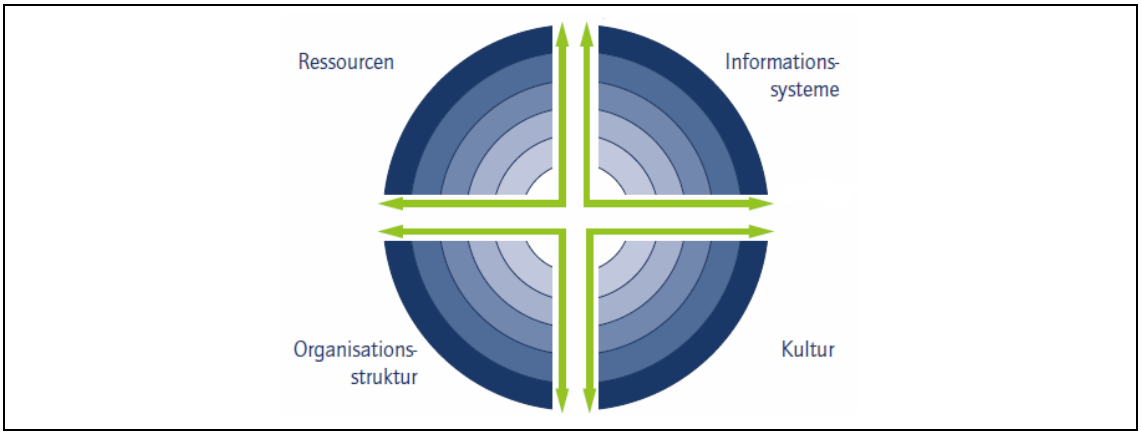

Abbildung 6: Bewertungsschema Industrie 4.0 Maturity Index[118]

Die Prinzipien haben nicht gleich viele Fähigkeiten. So ist die *strukturierte Kommunikation* eines der Prinzipien des Gestaltungsfeldes *Ressourcen*. Hier wurden zwei Fähigkeiten – „effizient kommunizieren“ und „Schnittstellen aufgabengerecht gestalten“ – festgelegt.[119] Für das Gestaltungsfeld Kultur wurde die *Bereitschaft für Veränderung* als Prinzip identifiziert. Die Fähigkeiten hier sind „Fehler als Schätze“, „Offenheit für Innovationen“, „datenbasiertes Lernen und Entscheiden“, „fortlaufende Qualifikation“ und „Gestalten von Veränderung“.[120] Um die Fähigkeit beurteilen zu können werden entsprechende Fragen dazu gestellt. Die Antwortmöglichkeiten sind den sechs Reifegradstufen zugeordnet. Der Gesamtreifegrad in einem Gestaltungsfeld ergibt sich aus den Reifegraden der Antworten. Diese Herangehensweise erlaubt es auch

[117] Vgl. Schuh, G. et al. (2017), S. 19.
[118] Quelle: Schuh, G. et al. (2017), S. 19. (Ausschnitt)
[119] Vgl. Schuh, G. et al. (2017), S. 22.
[120] Vgl. Schuh, G. et al. (2017), S. 34.

später Handlungsempfehlungen bei der Weiterentwicklung zu geben.[121] Die Nutzenbewertung der Reifegradbewertung erfolgt über ein hierarchisches Kennzahlensystem.[122]

Die Mächtigkeit des Modells verteilt einzelne Betrachtungsfelder wie Daten und Schnittstellen auf unterschiedliche Untersuchungsgegenstände. Für eine fokussierte Betrachtung nur einer Dimension, wie Daten, Datenqualität und Datenanalytik, eignet sich das Modell daher nicht. Der datenanalytische Prozess ist auf mehrere Gestaltungsfelder verteilt und kann nicht zielorientiert verbessert werden.

Maturity Model for Industry 4.0 Readiness

Dieses Reifegradmodell legt einen Fokus auf die organisatorischen Aspekte von Industrie 4.0 und das Konzept von Smart Manufacturing[123]. Folgende Grundprobleme sollen mit dem Reifegradmodell behandelt werden:

- Industrie 4.0 wird von Unternehmen als hoch komplex angesehen.
- Industrie 4.0 wird von Unternehmen nicht verstanden und somit wird der Nutzen nicht erkannt.
- Unternehmen können ihren eigenen Status und die eigenen Fähigkeiten bezüglich Industrie 4.0 nicht einschätzen.

Das Modell identifiziert 62 Elemente die Industrie 4.0 ausmachen. Diese werden in neun Reifegradkategorien gegliedert. Die Reife in den Dimensionen wird mit einer fünfstufigen Reifegradskala gemessen. In Stufe eins sind keine Konzepte implementiert, die Industrie 4.0 ausmachen, während in Stufe fünf der Stand der Technik erreicht ist. Tabelle 7 gibt einen[124] Überblick über die Reifegradkategorien.

Für jedes der 62 Industrie 4.0-Elemente existiert eine Frage, die von Personen des Unternehmens auf einer 5-stelligen Likertskala beantwortet wird. Wobei „1“ dem geringsten Reifegrad und „5“ dem höchsten Reifegrad entspricht. Die Elemente werden gewichtet indem 23 Experten mit einem Fragebogen jedes auf einer Likertskala von 1 – nicht wichtig – bis 4 – sehr wichtig – bewerten. Aus dieser Bewertung wird der Durchschnittswert für jedes Element gebildet.[125] Für jede Dimension wird der Reifegrad anhand Formel 2-1 berechnet:

M…Reife
D…Dimension
I…Element
g…Gewichtungsfaktor
n…Anzahl der Elemente

$$M_D = \frac{\sum_{i=1}^{n} M_{DIi} * g_{DIi}}{\sum_{i=1}^{n} g_{DIi}} \qquad 2\text{-}1$$

[121] Vgl. Schuh, G. et al. (2017), S. 46 f.
[122] Vgl. Schuh, G. et al. (2017), S. 49 f.
[123] Smart Manufacturing beschreibt die intelligente Vernetzung in der Produktion. Ziel ist ein intelligentes Fertigungssystem. Vgl. Promotorengruppe Kommunikation der Forschungsunion Wirtschaft – Wissenschaft (2013), S. 71.
[124] Vgl. Schumacher, A. et al. (2016), S. 163 f.
[125] Vgl. Schumacher, A. et al. (2016), S. 164.

Tabelle 7: Reifegradkategorien und I 4.0 Elemente[126]

Reifegradkategorie	Industrie 4.0-Elemente
Strategie	Implementierungsroadmap Industrie 4.0, Verfügbare Ressourcen für die Implementierung, …
Führung	Bereitschaft der Vorgesetzten, Managementkompetenzen und –methoden,…
Kunden	Verarbeitung von Kundendaten, Digitalisierung im Verkauf und Service, …
Produkte	Individualisierung von Produkten, Digitalisierung von Produkten, Produktintegration in andere Systeme,…
Betrieb/Einsatz	Dezentralisierung von Prozessen, Modellierung und Simulation, Interdisziplinarität, …
Kultur	Wissensteilung, Wertschätzung von IKT[127] im Unternehmen, …
Menschen	IKT Kompetenz der Beschäftigten, Autonomie der Beschäftigten, …
Unternehmensführung	Schutz von geistigem Eigentum, Eignung von technologischen Standards, …
Technologie	Existenz von moderner IKT, Verwendung von mobilen Geräten, M2M[128] Kommunikation, …

Ein hierarchischer Aufbau ist in dem Reifegradmodell nicht zu erkennen. Es ist auch nicht bekannt wie die Kategorien abgeleitet wurden. Was jedoch besonders auffällt ist, dass die Datenperspektive nicht in einer eigenen Kategorie betrachtet wird was für die Wichtigkeit von Daten und datenanalytischer Prozesse für Industrie 4.0 eine große Schwäche des Reifegradmodells ist.

Reifegradmodell Industrie 4.0

Dieses Reifegradmodell misst den IST-Reifegrad eines Unternehmens in drei Dimensionen zu je zehn Reifegradstufen. Basierend auf dem IST-Zustand wird basierend auf der Strategie und der Ziele des Unternehmens ein SOLL-Zustand definiert.[129] Zur Erreichung des SOLL-Reifegrades werden durch das Reifegradmodell Handlungsempfehlungen gegeben. Die Schritte und die Unterstützung durch das Reifegradmodell sind in ein Projektvorgehen integriert.[130] Es werden zu Beginn Applikationsfelder definiert, in denen das Reifegradmodell eingesetzt werden soll. Für diese werden Träger identifiziert, die die wesentlichen Funktionen eines Applikationsfeldes darstellen. Die Produktion ist z. B. ein Applikationsfeld, welches durch den Einsatz des Reifegradmodells verbessert werden soll. Träger könnten die Produktionsplanung, das Rüsten und die Fertigung sein. Jeder Träger wird in den drei

[126] Quelle: in Anlehnung an Schumacher, A. et al. (2016), S. 164. (aus dem Englischen übersetzt)
[127] IKT: Informations- und Kommunikationstechnologie
[128] Maschine zu Maschine
[129] Vgl. Brunner, M. et al. (2016), S. 49.
[130] Vgl. Brunner, M. et al. (2016), S. 51.

Dimensionen bewertet für die IST-Zustandsfestlegung und danach wird ein SOLL-Zustand abgeleitet.[131]

Die drei Dimensionen sind *„Daten“*, „*Intelligenz“* und „*Digitale Transformation“*. Dimensionen sind in Kriterien und diese in Subkriterien unterteilt.[132] Tabelle 8 gibt eine Übersicht über die drei Dimensionen mit den dazugehörigen Kriterien und Subkriterien.

Tabelle 8: Reifegraddimensionen Reifegradmodell Industrie 4.0[133]

Dimension	Kriterium	Subkriterium
Daten	Big Data	Volume Velocity Variety Veracity Visualization
	Open-Ansätze	
	Security	
Intelligenz	Enabler	Identifizierbarkeit Lokalisierbarkeit Connectivity-Grad Speicherfähigkeit Sensorausstattung Ausstattung Aktoren Rechenfähigkeit
	Nutzung Intelligenz	Selbstständigkeit Reaktionsfähigkeit Anpassungsfähigkeit Kooperationsfähigkeit Mensch-Maschine Symbiose
Digitale Transformation	Mitarbeiter	Mitarbeiter (Können und Wollen) Führung (Dürfen)
	Transformation	Durchgehende digitale Modellbildung Simulation und Optimierung Ersetzen Materielles durch Digitale

[131] Vgl. Jodlbauer, H.; Schagerl, M. (2016), S. 1479.; Jodlbauer, H.; Schagerl, M. (2016), S. 1483.
[132] Vgl. Brunner, M. et al. (2016), S. 50.
[133] Quelle: eigene Darstellung in Anlehnung an Brunner, M. et al. (2016), S. 50.

Die Reifegradbewertung erfolgt auf der Ebene der Subkriterien mittels Referenztabellen. Für jedes Subkriterium ist eine 10-stufige Referenztabelle definiert. Die Reifegrade der Kriterien errechnen sich aus den Reifegraden der Subkriterien. Aus den Mittelwerten der Kriterien ergeben sich die Reifegrade der Kriterien und aus diesen der Reifegrad der Organisation.[134] Die Gewichtung der Subkriterien erfolgt über die 10-stufigen Referenztabellen. Nicht jeder Stufe ist eine Eigenschaft zugeordnet. Wichtige Eigenschaften werden auf eine höhere Stufe gestellt. Es können Stufen auch leer bleiben, um Gewichtungen noch mehr zu prononcieren.[135]

In der Dimension *Daten* wird die Fähigkeit der Organisation bewertet große Datenmengen aus unterschiedlichen Quellen schnell zu verarbeiten. Diese Bewertung erfolgt mit den Aspekten von Big Data. Das Kriterium „Open-Ansätze“ trifft eine Aussage für die Reife der Vernetzung zum Datenaustausch. Folglich existiert das Kriterium „Security“, welches die Informationssicherheit bezüglich von Manipulation von Daten und unerlaubtem Zugriff bewertet.[136]

Die Dimension *Intelligenz* bewertet die Fähigkeit von einer Organisation maschinell und autonom zu handeln. Das Kriterium „Enabler“ stellt die Hardwareausrüstung unter genauere Betrachtung, während das Kriterium „Nutzung maschineller Intelligenz“ bewertet wie gut Systeme mit Menschen oder anderen Maschinen selbständig interagieren.[137]

Die *digitale Transformation* ist die dritte Dimension im Reifegradmodell. Hier wird bewertet wieviel Wertschöpfung in der digitalen Welt erreicht werden kann. Um die digitale Transformation vollziehen zu können, ist es nötig die Mitarbeiter entsprechend gut auszubilden und zu führen. Daher wird in dieser Dimension das Mitarbeiterkriterium bewertet.[138]

Das Reifegradmodell ist hierarchisch aufgebaut und erlaubt dadurch die Analyse des IST-Zustandes der Organisation hinsichtlich von Industrie 4.0. Es existieren wichtige Subkriterien, die für eine ganzheitliche Analyse eines datenanalytischen Prozesses notwendig sind. Diese werden jedoch nicht verknüpft, um diesen wesentlichen Prozess genauer zu analysieren.

Reifegradbasierte Integration von Industrie 4.0

Das Reifegradmodell ist Bestandteil eines Vorgehens, welches Unternehmen bei der Implementierung von Industrie 4.0 Methoden unterstützen soll. Der Reifegrad wird in fünf Kategorien zu je vier Stufen gemessen. Die Erhebung erfolgt mittels Fragen zu den Kategorien und haben den Einsatz von Industrie 4.0 Methoden zum Inhalt. Jeder Methode ist ein Referenzreifegradprofil hinterlegt. Wird ein bestimmtes IST-Profil festgelegt, und in der Potenzialfestlegung, welche nach der IST-Zustandsbestimmung folgt, wird festgestellt, dass eine bestimmte Methode im Unternehmen implementiert werden soll, so kann festgestellt werden, wo der Reifegrad verbessert werden muss, um die Methode einzuführen. Die Methoden werden einer vordefinierten Toolbox

[134] Vgl. Brunner, M. et al. (2016), S. 50.
[135] Vgl. Jodlbauer, H.; Schagerl, M. (2016), S.1474f.; Jodlbauer, H.; Schagerl, M. (2016), S.1478f.
[136] Vgl. Jodlbauer, H.; Schagerl, M. (2016), S. 1475 ff.
[137] Vgl. Jodlbauer, H.; Schagerl, M. (2016), S. 1477.
[138] Vgl. Jodlbauer, H.; Schagerl, M. (2016), S. 1478.

entnommen. Generell handelt es sich bei den Industrie 4.0 Methoden um Methoden der Produktions- und Logistikoptimierung, die durch IKT weiterentwickelt werden.[139]

Der Fokus der Bewertung wird u. a. auf der Kompetenzentwicklung der Mitarbeiter, einer Risiko- und Potenzialabschätzung für die Methodeneinführung und auf dem Umgang mit Daten gelegt. Die Datenerfassung, Weiterverarbeitung und Nutzbarmachung ist eine der Untersuchungsfelder des Vorgehens. Daten bilden jedoch keinen Schwerpunkt, wenngleich Daten als eine Grundvoraussetzung für die Weiterentwicklung der Methoden zu Industrie 4.0 gesehen werden.[140] Die Verbesserung durch die Reifegraderhöhung wird durch die Kennzahlen, die im Einsatzbereich existieren gemessen.[141]

Das Reifegradmodell hat zum Ziel die Organisation zu verbessern und entspricht somit der in dieser Arbeit gültigen Definition eines Reifegradmodells. Es werden jedoch keine spezifischen Methoden oder Kategorien für die Datenanalyse definiert.

Data Quality Assessment of Maintenance Reporting Procedures

Dieses Assessment unterscheidet sich von den bisher vorgestellten Reifegradmodellen, da es nicht den klassischen Ansatz eines Reifegradmodells anwendet. Es wurden keine Kategorien und Reifegradstufen definiert. Trotzdem wird eine IST-Analyse (assessment) durchgeführt, welche die Datenqualität von Instandhaltungsrückmeldungen bewerten soll. Da der Bedarf für dieses Assessment von den großen Datenmengen und Anforderungen durch Industrie 4.0 abgeleitet wurde, wird es den Reifegradmodellen der Digitalisierung zugeordnet. Tabelle 9 gibt einen Überblick über die Bewertungskategorien, deren Ziel und die zugeordnete Datenqualitätsdimension.

Tabelle 9: Assessmentkategorien nach Madhikermi[142]

Datenqualitäts-konzept	Untersuchungsgegenstand	Datenqualitäts-dimension
Semantische Qualität (semantic quality)	Wie sehr kann den Rückmeldungen getraut werden?	Glaubwürdigkeit
Sprachliche Qualität (language quality)	Sind die Rückmeldungen vollständig, bzw. ausreichend detailliert ausgeführt?	Vollständigkeit
Wissensqualität (knowledge quality)	Sind die Rückmeldungen auf dem aktuellen Stand? Wie groß ist die Zeitspanne zwischen der Durchführung und der Rückmeldung?	Aktualität

Ziel des Assessments ist es, die Qualität von Instandhaltungsmeldungen dynamisch festzustellen.[143] Das Assessment wird auf einem generischen Rahmenwerk[144] zur Beschreibung von Datenqualität aufgebaut. Aus diesem werden drei Konzepte der Datenqualität gewählt, die den industriellen Bedarf am besten abdecken.

[139] Vgl. Lanza, G. et al. (2016), S. 76 f.
[140] Vgl. Lanza, G. et al. (2016), S. 77 f.
[141] Vgl. Lanza, G. et al. (2016), S. 79.
[142] Quelle: Eigene Darstellung in Anlehung an Madhikermi, M. et al. (2016), S. 150.
[143] Vgl. Madhikermi, M. et al. (2016), S. 146.
[144] Krogstie's data quality framework (siehe: Krogstie, J. et al. (1995))

Zu den Datenqualitätsdimensionen wurden Metriken definiert, die die Fragestellung des Untersuchungsgegenstandes quantitativ bewerten sollen. Die Datenqualitätsdimensionen werden als Bewertungskategorien verwendet, die untereinander als gleichgewichtet definiert wurden. Das Assessment wurde bei einem OEM angewandt, der mit 54 Standorten Instandhaltungsserviceleistungen bei seinen Produkten anbietet. Über die drei Datenqualitätsdimensionen wurden die Instandhaltungsreports eines jeden Standortes bewertet und diese untereinander gereiht. Das Vorgehen wird verwendet, um die Rückmeldequalität und die Servicequalität zu verbessern und Kosten zu reduzieren.[145]

Dieses Modell bewertet die Reife der Daten mittels eines Auszuges der Datenqualitätsdimensionen und entspricht folglich dem Anspruch, den Prozess qualitativ zu verbessern. Es fehlt jedoch am hierarchischen Aufbau und dem Fokus auf den datenanalytischen Prozess, der die Daten verarbeitet.

Vergleich der Reifegradmodelle bezüglich der Datenperspektive

Die ausgewählten Reifegradmodelle behandeln weitestgehend Industrie 4.0 und damit verbunden die Datenperspektive. Die Bewertung von Daten bezogen auf den generischen Datenanalyseprozess wird in keinem der Modelle durchgeführt.

Der „Industrie 4.0 Maturity Index“ legt den Fokus auf eine ganzheitliche Umsetzung von Industrie 4.0. Daten sind dafür ein Mittel zum Zweck, die in unterschiedlichen Perspektiven des Reifegradmodells unter verschiedenen Blickwinkeln betrachtet werden. Die Verarbeitung der Daten durch den Analyseprozess wird nicht betrachtet.

Das „Maturity Model for Industry 4.0 Readiness“ betrachtet ausschließlich Kundendaten direkt. Ansonsten finden Daten nur in der Technologieebene Beachtung, indem bewertet wird, wie modern IK-Technologien eingesetzt werden. Wie die Daten für Analysen beschaffen sein müssen findet keine Beachtung, obwohl sich die Reifekategorie Betrieb/Einsatz mit der Umsetzung der Industrie 4.0 Konzepte befasst.

Das „Reifegradmodell Industrie 4.0“ behandelt Daten mehr in der Tiefe. Der Fokus liegt auf Big Data und den Eigenschaften von Big Data. Diese Eigenschaften sind jedoch inhärente Dimensionen der Datenqualität. Es finden auch die Enabler für die Erfassung von Daten und die Umsetzung der Ergebnisse Beachtung. Es fehlt jedoch der Konnex zum Analyseprozess und den Use-Cases. Die erfolgreiche Umsetzung datenanalytischer Projekte kann damit nicht abgeschätzt werden.

Das Modell der „Reifegradbasierten Integration von Industrie 4.0“ legt den Fokus auf die Implementierung. Die Datenperspektive wird hauptsächlich unter dem Blickwinkel der Erfassung betrachtet. Die Datenauswertung findet sich unter dem Schlagwort der Datenkorrelation. Eine genauere Betrachtung der analytischen Anforderungen erfolgt nicht.

Das „Data Quality Assessment of Maintenance Reporting Procedures“ betrachtet als einziges vorgestelltes Modell einen Auszug der Datenqualitätsdimensionen. Es wird ein generisches Rahmenwerk verwendet, aus welchem die relevanten Untersuchungsgegenstände abgeleitet und mit den Datenqualitätsdimensionen

[145] Vgl. Madhikermi, M. et al. (2016), S. 150 ff.

gemessen werden. Es fehlt jedoch der Bezug auf den datenanalytischen Prozess und die Betrachtung weiterer Datenqualitätsdimensionen.

Jedes der vorgestellten Modelle hat bezogen auf die Datenperspektive Stärken und Schwächen. Keines betrachtet die Anforderungen, die an Daten gestellt werden aus dem Blickwinkel des verarbeitenden Prozesses, welcher in diesem Fall der Datenanalytische ist. Des Weiteren fehlt es an der Objektivierung und der Sicherstellung der Vergleichbarkeit der Anforderungen. Dafür bietet sich die Datenqualitätsbetrachtung an. Deren Konzepte, die bisher hauptsächlich in der BI[146] Anwendung finden, müssen auf industrielle Anwendungen ausgedehnt werden.

2.4 Relevanz für die Arbeit

In diesem Kapitel wurde die Modelltheorie beschrieben, welche ein wichtiger Input für die Entwicklung eines Reifegradmodells darstellt. Der wesentliche Inhalt ist die Festlegung der Definition eines Reifegeradmodells, welche für diese Arbeit gilt. Aufbauend darauf wird das Reifegradmodell so entworfen, dass es den datenanalytischen Prozess anhand von Kategorien analysiert und bewertet. Der Prozess soll innerhalb dieser Kategorien verbesserbar sein, um den Reifegrad zu erhöhen. Durch den hierarchischen Aufbau soll eine kontinuierliche Verbesserung möglich sein, welche den datenanalytischen Prozess erlaubt effizienter ausgeführt zu werden.

Des Weiteren wurde der Aufbau und folglich das Vorgehen für die Entwicklung des Reifegradmodells definiert. Das Modell findet seine Grundlagen und Legitimation in einer theoretischen Fundierung. Die empirische Ebene bildet die Schnittstelle zur Realwelt, welche über die Bewertungsebene in den Reifegradkategorien abgebildet wird. Dieser hierarchische Aufbau erlaubt eine aggregierte Bewertung von Subprozessschritten der Datenanalyse und eine Dekomposition von Handlungsempfehlungen.

Am Ende des Kapitels wurden Reifegradmodelle gezeigt, welche Organisationen im Fokus der Digitalisierung und Industrie 4.0 betrachten, bzw. die Datenperspektive allein bewerten. Trotz des großen Umfangs einiger Modelle oder des spezifischen Fokus auf die Daten, gibt es nach aktuellem Stand dieser Arbeit kein Reifegradmodell, welches die Organisation anhand seiner Fähigkeit bewertet einen datenanalytischen Prozess anhand eines definierten Vorgehens durchzuführen.

Da das entwickelte Reifegradmodell den datenanalytischen Prozess bewerten soll, wird in den folgenden beiden Kapiteln näher auf das Daten- und Informationsmanagement mit dem Schwerpunkt der Datenqualität sowie auf Grundlagen zur Datenanalyse eingegangen.

[146] BI: Business Intelligence. Beschreibt die Analyse von Daten zum Zweck der Berichtserstellung mittels einfacher Werkzeuge wie Abfragen und Berichtssystemen über Datenvisualisierungen bis hin zu komplexem Data-Mining. Der Fokus liegt auf strategischen Auswertungen von Kundenverhalten oder Absatzzahlen und Ähnlichem. (Vgl. Dippold, R. et al. (2005), S. 9 & 281.; Krcmar, H. (2015), S. 152 f.) Big Data Analytics auf operative Ebene (industrielle Anwendung) wird nicht dazugezählt.

3 Daten- und Informationsmanagement

Daten und Informationen gewinnen durch die vierte industrielle Revolution als zusätzlicher Produktionsfaktor verstärkt an Bedeutung. Das Daten- und Informationsmanagement umfasst die Methoden und Aufgaben diesen Produktionsfaktor zu planen und zu steuern.

In diesem folgenden Kapitel werden die Begriffe Daten, Informationen und Wissen definiert und abgegrenzt. Darauf aufbauend wird das Daten- und Informationsmanagement beschrieben, wobei das Wissensmanagement bewusst nicht behandelt wird, welches die Planung, Verwertung und Weiterentwicklung von Wissen beschreibt.[147] Der Schwerpunkt dieses Kapitels liegt auf dem Datenqualitätsmanagement.

3.1 Begriffsdefinition Daten, Informationen und Wissen

Daten und deren Verarbeitung bilden die Grundlage für die Weiterentwicklung der Wirtschaft von Industrie 3.X zu Industrie 4.0. Wird von Industrie 4.0 gesprochen, so wird unweigerlich von der Digitalisierung und deren Nutzung zur Datengewinnung und -weiterverarbeitung gesprochen. Die Etablierung von Daten als treibender Produktionsfaktor schafft die Voraussetzungen neue Potenziale der Wertschöpfung zu heben.

In Verbindung mit Industrie 4.0 fällt häufig der Begriff Big Data. Er wurde das erste Mal von MASHEY in den späten 1990er Jahren erwähnt und von DIEBOLD[148] im Jahr 2000 zum ersten Mal in einem wissenschaftlichen Paper verwendet. *Big* bezog sich dabei nur auf die Datenmenge. Gartner[149] formulierte 2001 drei Dimensionen von Big Data, welche bis heute als die drei V‘s – *volume* (Menge), *velocity* (Geschwindigkeit) und *variety* (Vielfältigkeit) – bekannt sind.[150] Die wahllose Aufzeichnung der Daten aus unterschiedlichen Datenquellen, ließ die Datenqualität sinken, was zur Folge hatte, dass sich mit *veracity* (Glaubwürdigkeit) eine weitere Dimension von Big Data etablierte. Der Wert der Daten als fünfte Dimension *value* hat sich in den letzten Jahren etabliert.[151] Abbildung 7 gibt einen Überblick über die relevanten fünf V‘s von Big Data mit einer kurzen Beschreibung ihrer Bedeutung. Die genaue Definition der zugrundeliegenden Ressource Daten und deren weiterverarbeiteten Produkte ist in diesem Zusammenhang umso wichtiger.

[147] Vgl. Bodendorf, F. (2006), S. 133 f.
[148] Siehe Diebold, F. X. (2003)
[149] Frühere META Group
[150] Vgl. Dean, J. (2014), S. 3 f.
[151] Vgl. Bernerstätter, R. (2018c), S. 36 f.

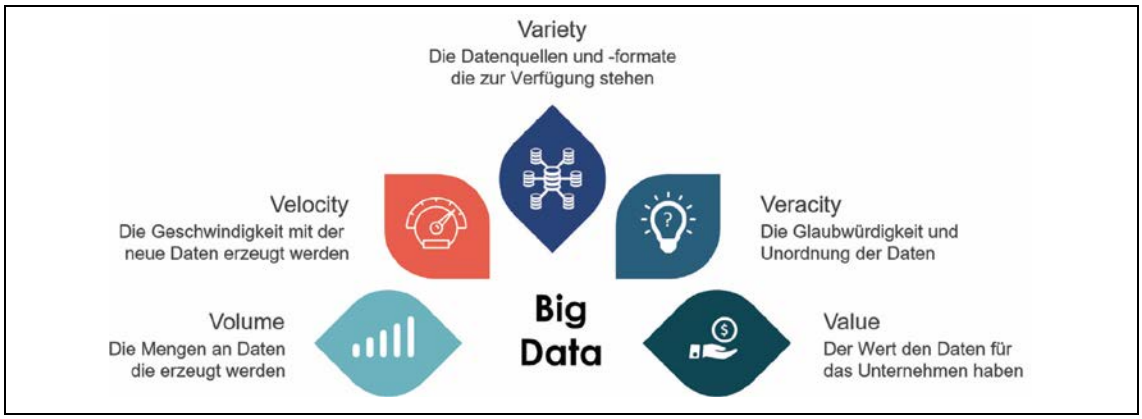

Abbildung 7: Dimensionen von Big Data[152]

Abbildung 8 zeigt mit der Wissenspyramide einen Versuch der Abgrenzung von Zeichen, Daten, Informationen und Wissen. Die Abgrenzung zwischen Daten und Information erfolgt über den Kontext, in dem die Daten gesehen werden müssen. Der Name Müller wird über den Kontext als Nachname verwendet und nicht als Beruf. Noch eindeutiger wird der Zusammenhang bei Zahlen. So ist beispielsweise „2,54“ ohne weiter Angaben relativ aussagelos, 2,54 cm setzt die Zahl jedoch in Kontext eines Längenmaßes. Mit der Vernetzung mit anderen Informationen und Erfahrung entsteht das Wissen, dass 2,54 cm einem Zoll entsprechen.[153]

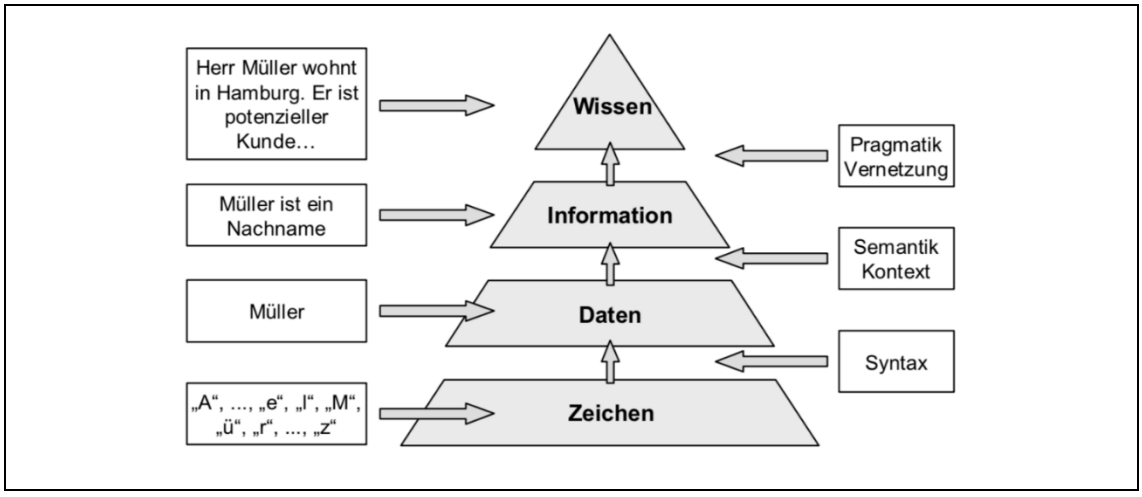

Abbildung 8: Wissenspyramide[154]

NORTH entwickelte die Wissenspyramide zur Wissenstreppe (Abbildung 9) weiter. Dabei sind die Definitionen von Zeichen, Daten und Informationen und deren Übergänge ineinander unverändert zur Wissenspyramide. Auch das Wissen ist als kontextabhängige Verknüpfung von Informationen definiert.[155] Eine ausführlichere Definition von Wissen lautet wie folgt:

[152] Quelle: Bernerstätter, R. (2018c), S. 37.
[153] Vgl. North, K. (2011), S. 36 f.; Otto, B.; Österle, H. (2016), S. 27.; Otto, B.; Österle, H. (2016), S. 196 ff.
[154] Quelle: Bodendorf, F. (2006), S. 1.
[155] Vgl. North, K. (2011), S. 36 f.

„[…] Wissen als die Gesamtheit der Kenntnisse und Fähigkeiten, die Personen zur Lösung von Problemen einsetzen. Dies umfasst sowohl theoretische Erkenntnisse als auch praktische Alltagsregeln und Handlungsanweisungen. Wissen stützt sich auf Daten und Informationen; ist im Gegensatz zu diesen jedoch immer an Personen gebunden. Wissen entsteht als individueller Prozess in einem spezifischen Kontext und manifestiert sich in Handlungen."[156]

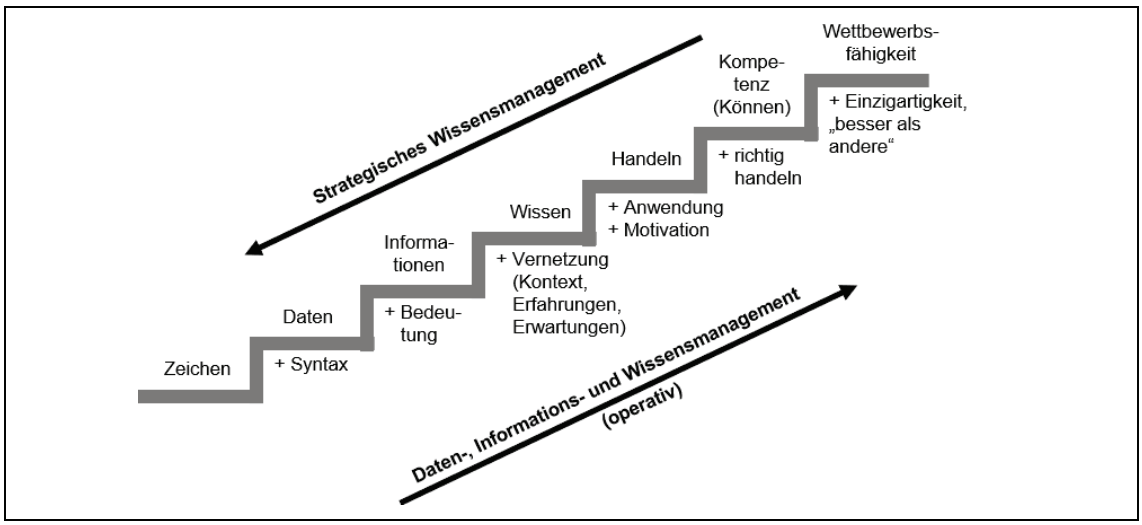

Abbildung 9: Wissenstreppe[157]

Wissen ist speziell von Wert, wenn es eine Organisation befähigt Handlungen zu setzen. NORTH sieht den Zusammenhang unter dem Aspekt der Mitarbeiterqualifikation. Mitarbeiter müssen befähigt und motiviert werden, das vorhandene Wissen einzusetzen.[158]

Die Kompetenz zeigt sich, wenn Wissen angewandt wird. Die Handlung ist eine undifferenzierte Anwendung von Wissen und Empfehlungen. Kompetenz ist, wenn die Handlung hinterfragt wird und richtig und der Situation entsprechend gehandelt wird. Hier zeigt sich die Notwendigkeit der Erfahrung, um differenziert Wissen anzuwenden.[159]

Data-Mining kann kompetentes Handeln durch die richtige Auswahl von Handlungsempfehlungen unterstützen. Die Basis dafür wird bereits in den ersten Phasen des Data-Mining Prozesses gelegt in der die Ziele und die Zusammenhänge im Unternehmen definiert werden.

Metadaten:

Metadaten sind eine Sonderform von Daten, die weder in der Wissenspyramide noch in der Wissenstreppe einordenbar sind. Bei Metadaten handelt es sich um Daten über Daten. Es sind sogenannte beschreibende oder systemtechnische Daten.[160] Sie

[156] North, K. (2011), S. 37.
[157] Quelle: in Anlehnung an North, K. (2011), S. 36.
[158] Vgl. North, K. (2011), S. 38.
[159] Vgl. North, K. (2011), S. 38.
[160] Vgl. Dippold, R. et al. (2005), S. 97 f.

beschreiben neben den Formaten, der Codierung oder der Einheit, auch die Herkunft und die Aktionen, die im Laufe des Lebenszyklus der Daten durchgeführt wurden.[161]

Datenanalytisch ist dieser Aspekt ebenfalls wichtig, da Mitarbeiter das Wissen, welches durch Modelle erzeugt wird, anwenden sollen. Dazu können Handlungsempfehlungen gegeben werden, wie die Umsetzung erfolgen soll.

Stammdaten:

Stammdaten sind faktisch unveränderliche Basisdaten eines Unternehmens, wie Produktspezifikationen, Konstruktionsdaten oder Kundendaten.[162]

Bewegungsdaten:

Bewegungsdaten verändern sich mit der Zeit und werden auch Transaktionen genannt. Sie werden laufend erzeugt, wie Sensormessungen oder Kundenaufträge. Sie sind im Vergleich zu Stammdaten einer häufigen Veränderung unterworfen und liegen in größerer Menge vor.[163]

Ein Mehrwert an Wettbewerbsfähigkeit ergibt sich, wenn das Wissen neue Kernkompetenzen generiert. Kernkompetenzen differenzieren ein Unternehmen vom Mitbewerber und generieren für den Kunden einen Wert.[164]

Klassische Methoden des Ressourcenmanagements sind auf Daten und Informationen nicht unreflektiert übertragbar. REDMAN hat dafür sechs Gründe beschrieben:[165]

- Daten sind abstrakt. Von Daten werden nur unterschiedliche Darstellungen und Repräsentationen gesehen. Sie sind folglich nicht konkret.
- Daten sind beliebig vervielfältigbar und übertragbar.
- Daten werden nicht verbraucht. Der Nutzen kann sich über ihren Lebenszyklus hinweg verändern.
- Daten sind nicht austauschbar, da Mehrwert aus dem Vergleich der Daten entsteht. Neue oder andere Daten verändern die Aussage.
- Bis auf wenige Ausnahmen, wie Kundendaten, gibt es für Daten keinen Käufer- oder Verkäufermarkt. Es gibt daher keinen Preis wodurch sich der Wert schwerer bestimmen lässt.
- Daten sind dynamisch.

Um die Wettbewerbsfähigkeit zu garantieren bedarf es einer kompetenten Umsetzung des Wissens. Dieses Wissen muss jedoch auf belastbaren Informationen und Daten beruhen. Dafür müssen die Daten und Informationen richtig erfasst, verarbeitet und übertragen werden und eine hohe Qualität vorweisen. Daten- und Informationsmanagement unterstützen diese Anforderungen. Folglich werden die Themen Daten- und Informationsmanagement und Datenqualität erörtert.

[161] Vgl. Bodendorf, F. (2006), S. 39.
[162] Vgl. Otto, B.; Hinderer, H. (2009), S. 23.
[163] Vgl. Otto, B.; Hinderer, H. (2009), S. 23.
[164] Vgl. North, K. (2011), S. 38 f.
[165] Vgl. Redman, T. C. (1996) zitiert nach; Dippold, R. et al. (2005), S. 245.

3.2 Abgrenzung Daten- und Informationsmanagement

Wie die Abgrenzung zwischen Daten und Informationen, so ist auch die Abgrenzung zwischen Daten- und Informationsmanagement schwierig. Das Datenmanagement umfasst alle Tätigkeiten zur Bereitstellung der Ressource Daten in der richtigen Menge und Qualität für den informationsverarbeitenden Prozess, der aus Daten über den Informationszwischenschritt das Endprodukt Wissen herstellt, welches als Entscheidungsunterstützung gilt.

3.2.1 Datenmanagement

DIPPOLD definiert das Datenmanagement wie folgt:

„Wir verstehen unter Datenmanagement einerseits alle Prozesse, welche der Planung, Beschaffung, Organisation, Nutzung und Entsorgung der Unternehmensressource Daten dienen, und andererseits die Organisationseinheiten, welche für diese Prozesse gesamthaft verantwortlich sind.“[166]

Abbildung 10 zeigt die Bereiche, die die Ressource Daten ausmachen und durch ein durchgängiges Datenmanagement berücksichtigt werden müssen. Das Management des Lebenszyklus der Daten nimmt in der obigen Definition eine spezielle Rolle ein. In dieser Arbeit wird der Fokus auf die Qualität der Daten gelegt.

Das Datenmanagement umfasst die Bereiche der Datenbankadministration, der Datenadministration und der Datenarchitektur bzw. Datenmodellierung. Diese Bereiche müssen für ein erfolgreiches Datenmanagement durchgehend umgesetzt werden. Über die letzten Jahre haben sich speziell in der Datenarchitektur Neuerungen ergeben, die sich auch auf die Datenqualität auswirken.[167]

Die Datenmanagementstrategie definiert, wie die von DIPPOLD beschriebenen Aufgaben des Datenmanagements erfolgen sollen. Ein spezieller Faktor umfasst, dass die Daten und die Prozesse, die sich mit ihnen befassen miteinander abgestimmt werden müssen. Dieser Faktor gewinnt verstärkt an Bedeutung, da Daten in immer mehr Formaten vorkommen. Strukturierte und unstrukturierte Daten werden von unterschiedlichen Quellen erzeugt und müssen in einem gemeinsamen System erfasst und verarbeitet werden.[168]

Im Gegensatz zu strukturierten Daten liegen unstrukturierte Daten nicht in einem Format vor die es erlauben sie in Zeilen und Spalten in einer Datenbank abzubilden. Dazu zählen Textdateien, Ton- und Videoaufzeichnungen oder Bildmaterial.[169]

Eine durchgängig geplante Datenarchitektur hilft dabei die unterschiedlichen Datenformate besser verwalten zu können. Sie bestimmt die eingesetzte Technologie, die verwendeten Daten und deren Darstellung und die Zugänglichkeit für Benutzer und Systeme. Die Datenarchitektur hat durch die Festlegung von Regeln in einer Phase mit konzipierendem Charakter einen großen Einfluss auf die Datenqualität. Häufige

[166] Dippold, R. et al. (2005), S. 21.
[167] Vgl. Dippold, R. et al. (2005), S. 22.
[168] Vgl. Krcmar, H. (2015), S. 178 f.
[169] Vgl. Apel, D. et al. (2015), S. 99 f.

Transformationen bei schlecht geplanten Systemen führen zu Feldfehlern, die sich auf die Codierung der Daten auswirken. [170]

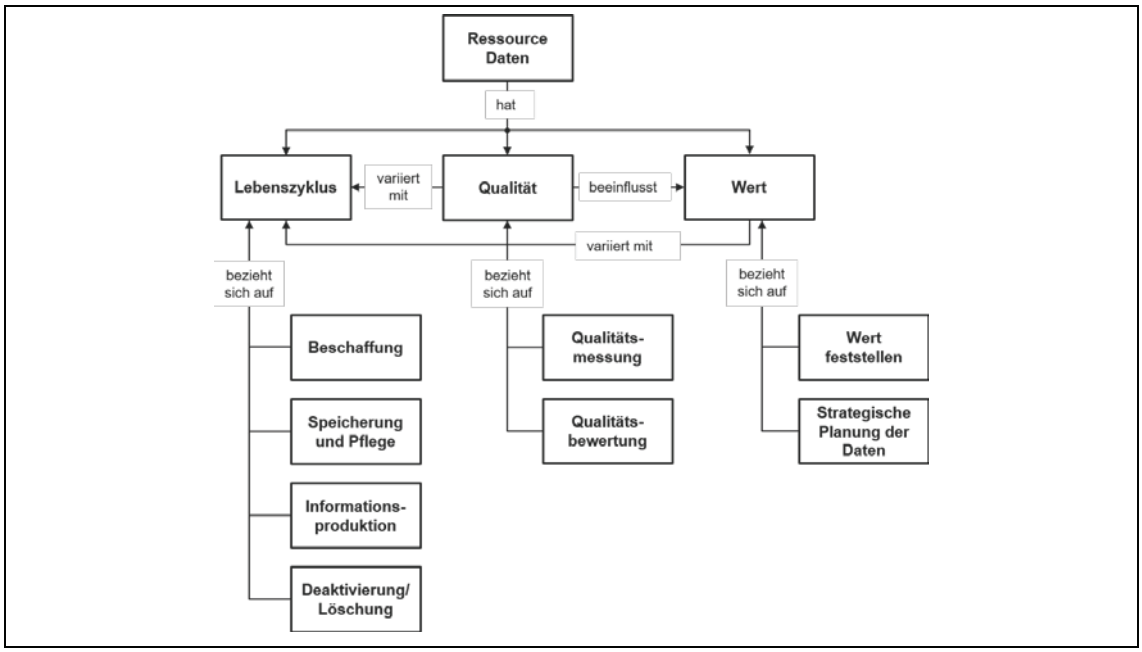

Abbildung 10: Managementbereiche der Ressource Daten[171]

Während sich die Architektur mit dem hierarchischen Aufbau des Datenflusses beschäftigt, beschreibt die Datenmodellierung, wie Unternehmensdaten aufgebaut sein müssen. Das Datenmodell definiert wie die Daten in den Systemen gespeichert werden. Die genaue Beschreibung umfasst die Beziehungen unter Datentabellen, wie die Modifizierung und Transformation erfolgt und welche Operationen angewandt werden können. Die Art der Modellierung orientiert sich an den Indikatoren wie Preis, Verbreitung, Leistungsfähigkeit oder Integrationsmöglichkeiten. Am weitesten verbreitet ist das relationale Datenmodell.[172]

Die Datenbank- und Datenadministration haben zur Aufgabe die Vorgaben der Architektur und Modellierung umzusetzen und zu überwachen.[173]

Die stetig wachsende Menge an Daten und Datenquellen haben neue Entwicklungen im Datenmanagement gebracht. So gehen einfache Datenbanksysteme verstärkt in Data-Warehouses auf oder die Sammlung von Daten erfolgt in Data Lakes. Um die Verteilung von Daten sicherzustellen haben sich auch Data-Marts etabliert. Jedes dieser Konzepte hat Stärken und Schwächen, die in der Konzeptionsphase beachtet werden müssen.

Datenbanken sind eine Sammlung von strukturierten Daten, die meist in Tabellenform dargestellt sind. Die Tabellen stehen untereinander in Beziehung.[174] Das Datenbanksystem erlaubt die Organisation, Erstellung und Veränderung der

[170] Vgl. Apel, D. et al. (2015), S. 30 f.
[171] Quelle: in Anlehnung an Otto, B. (2015), S. 238. (aus dem Englischen übersetzt)
[172] Vgl. Krcmar, H. (2015), S. 70 f.
[173] Vgl. Dippold, R. et al. (2005), S. 22.
[174] Vgl. Lusti, M. (1997), S. 315.

Datenbanken und das Datenbankmanagementsystem bildet die Schnittstelle zum Benutzer, um über das Datenbanksystem mit den Datenbanken zu interagieren.[175]

Datenbanken sind ein alt bekanntes Konzept. Etwas neuer sind Data-Warehouses. Ein Data-Warehouse umfasst mehrere Datenbanken, die Daten aus unterschiedlichen Quellen strukturiert abbilden. Während eine Datenbank in den operativen Geschäftsbetrieb stark integriert ist, ist ein Data-Warehouse davon losgelöst. Es greift auf die Datenbanken zu und hält einen redundanten Datenbestand der Datenbanken vor. Das Data-Warehouse ist jedoch für analytische Aufgabenstellungen optimiert. Seine Architektur und Datenmodellierung müssen so gestaltet sein, dass analytische Prozesse die Daten aus unterschiedlichen Quellen benötigen schnell durchgeführt werden können. Die Ergebnisse der Abfragen werden an die unterschiedlichen Stellen in einem Unternehmen zentral aus dem Data-Warehouse verteilt. Im Fokus steht der Prozess des Data-Warehousing, der eine benutzergerechte Datenbereitstellung zur Informationsgewinnung zum Ziel hat. Ein Data-Warehouse soll für alle Anfragen in einem Unternehmen den Single Point of Truth bilden.[176]

Die Data-Marts sind Teildatenmengen von Data-Warehouses, die für spezielle Unternehmensanwendungen – Einkauf, Vertrieb, Personalwesen, usw. – zur Verfügung gestellt werden. Sie treffen die Anforderungen dieser Abteilungen und stellten die geforderten Daten und Abfragen, durch das geringere Volumen schnell zu Verfügung.[177] Abbildung 11 gibt einen Überblick über die Zusammenhänge der einzelnen Datenspeicherkonzepte, mit dem Data-Warehouse im Zentrum.

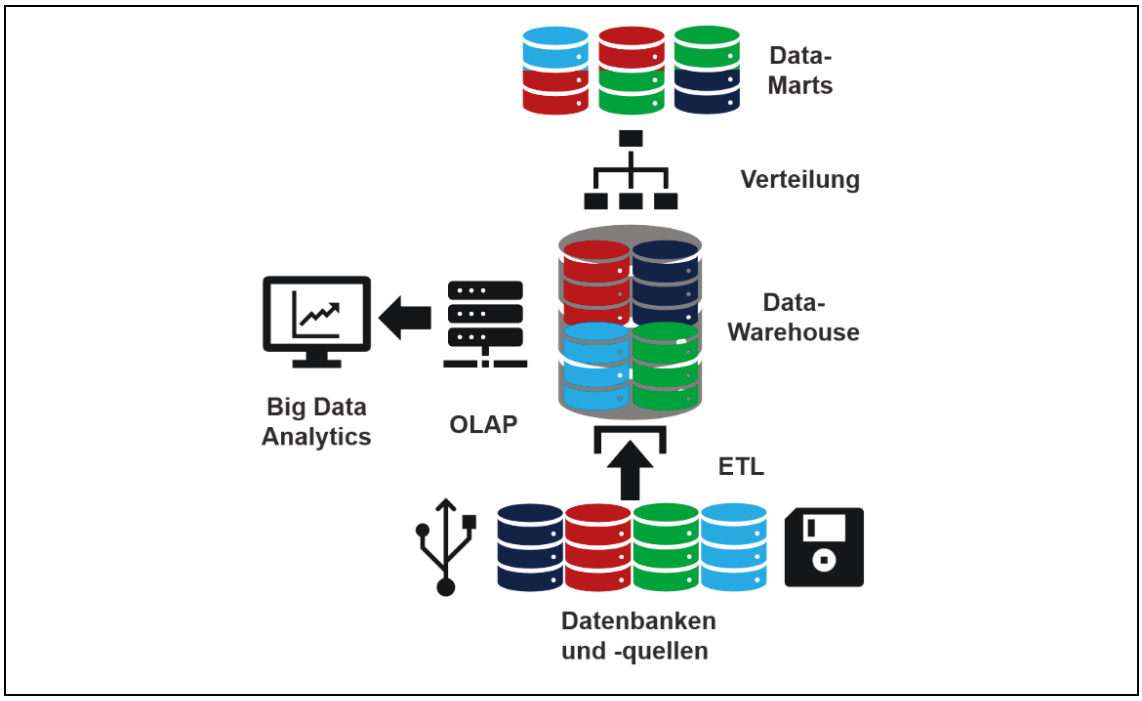

Abbildung 11: Übersicht Datenspeicherkonzepte[178]

[175] Vgl. Voß, S.; Gutenschwager, K. (2001), S. 38.
[176] Vgl. Goeken, M. (2006), S. 16 & 22 f.
[177] Vgl. Bodendorf, F. (2006), S. 39.; Goeken, M. (2006), S. 31 f.
[178] Quelle: Eigene Darstellung

Das Konzept der Data Lakes entstand als Antwort auf die rasend wachsende Menge an Daten und Datenquellen und –formaten. Data Lakes umfassen eine Vielzahl dezentraler Datenquellen, welche Daten in strukturierter wie unstrukturierter Form und im Rohformat enthalten. Dies ist der wesentliche Unterschied zu einem Data-Warehouse, welches nur strukturierte Daten enthält. Die eigentliche Intention hinter Data Lakes ist die sehr einfache Implementierung, da keine besondere Datenarchitektur und –modellierung notwendig ist. Daten werden bei Bedarf, z. B. für Analysezwecke, entnommen und müssen erst dann beschrieben werden. Das Konzept hinter dieser einfachen Sammlung der Daten ist kostengünstig im Vergleich zu einem Data-Warehouse, da es einfach skalierbar ist. In den letzten Jahren haben sich jedoch kritische Stimmen gemehrt, die die einfache Handhabe der Datenquellen in einem Data Lake in Frage stellen. Es muss bedacht werden, dass die klassische Datenbearbeitung wie in einem ETL-Prozess[179], nur verschoben wird. Sobald die Daten benötigt werden, müssen sie transformiert und mit anderen Daten zu einem gemeinsamen Datenbestand abgeglichen werden, um Analysen darauf anzuwenden.[180] Ein Data Lake stellt somit – obwohl das neueste Konzept – nicht zwingend die beste Lösung für eine Datenbereitstellung für analytische Zwecke dar. Der Data Lake hat seine Berechtigung als Datenspeicher. Das Data-Warehouse ist aufgrund seiner Architektur für die Bereitstellung von Daten für Analysezwecke zu bevorzugen und bildet eine Schnittstelle zwischen einem Data Lake und dem Endbenutzer.

„Unter Cloud Computing versteht man ein IT-basiertes Bereitstellungsmodell, bei dem Ressourcen sowohl in Form von Infrastruktur als auch Anwendungen und Daten als verteilter Dienst über das Internet durch einen oder mehrere Leistungserbringer bereitgestellt wird.“[181] Cloudservices haben speziell klein- und mittelständische Unternehmen befähigt Big Data Initiativen zu setzen.[182] Neben Anwendungen und Plattformen kann auch Speicherplatz in die Cloud ausgelagert werden. Abhängig vom Servicepaket könne die Daten nur gespeichert werden oder auch bereits gesäubert und vorverarbeitet für die folgende Anwendungen.[183]

Es zeigt sich, dass sich das Datenmanagement von der anfänglichen einfachen Modellierung von Daten und der Planung von Datenbanken zu einer äußerst komplexen Disziplin entwickelt hat. Die Kernelemente der Administration, Modellierung und Planung bleiben, es haben sich jedoch die Möglichkeiten und Anforderungen weiterentwickelt. Ein gut geplantes Data-Warehouse ist für datenanalytische Prozesse am besten geeignet. Es fasst Daten aus unterschiedlichen Quellen zusammen, verteilt die Daten nach Bedarf beliebiger Verbraucher und erlaubt durch OLAP-Anwendungen[184] eine effiziente Vorverarbeitung für komplexe Analysen. Des Weiteren sollte es für das

[179] ETL: Extract, Transform and Load. Daten werden aus dem Datenbestand in das Data-Warehouse übertragen. Vgl. Bodendorf, F. (2006), S. 38.
[180] Vgl. Terrizzano, I. et al. (2015), S. 1 f.
[181] Krcmar, H. (2015), S. 723.
[182] Vgl. Schmidt, R.; Mohring, M. (2013), S. 136.
[183] Vgl. Schmidt, R.; Mohring, M. (2013), S. 139 f.; Krcmar, H. (2015), S. 725 f.
[184] OLAP: Online Analytical Processing. Ist ein in Data-Warehouses integrierter Prozess zur schnellen Datenaufbereitung, um schnell einfache Analyseergebnisse zur Verfügung stellen zu können. (vgl. Chamoni, P.; Gluchowski, P. (2006), S. 14 f.)

gesamte Unternehmen als zentraler Punkt der Datenbeschaffung, dem sogenannten Single Point of Truth, dienen.

3.2.2 Informationsmanagement

BRUST ET AL. definiert Informationsmanagement als „[…] konzeptionelle Maßnahmen, die eine systematische Informationsbereitstellung für die Geschäftsführung sicherstellen“.[185]

VOß UND GUTENSCHWAGER fassen auf zwei Seiten unterschiedliche Definitionen zum Informationsmanagement zusammen. Zwei davon sind:

„Informationsmanagement ist die wirtschaftliche Produktion, Verteilung und Nutzung des Produktionsfaktors Information für alle Bereiche und Ebenen betrieblicher Aufgabenbearbeitung.“[186]

„Aufgabe des Informationsmanagement ist es, dafür zu sorgen, [dass] Informationen effektiv (zielgerichtet) und effizient (wirtschaftlich) eingesetzt werden.“[187]

Für DIPPOLD „[umfasst das Informationsmanagement] aus der Sicht der Unternehmensführung das methodische Planen, Umsetzen und Controlling der betrieblichen Informationsversorgung. Darunter fallen alle konzeptionellen. technischen. methodischen und organisatorischen Maßnahmen, um eine hohe Qualität der Informationsbereitstellung im Unternehmen in inhaltlicher, zeitlicher und räumlicher Hinsicht sicherzustellen.“[188] Abbildung 12 zeigt den Prozess des Informationsmanagements. Die Daten und folglich das Datenmanagement bilden die Basis für das Informationsmanagement.

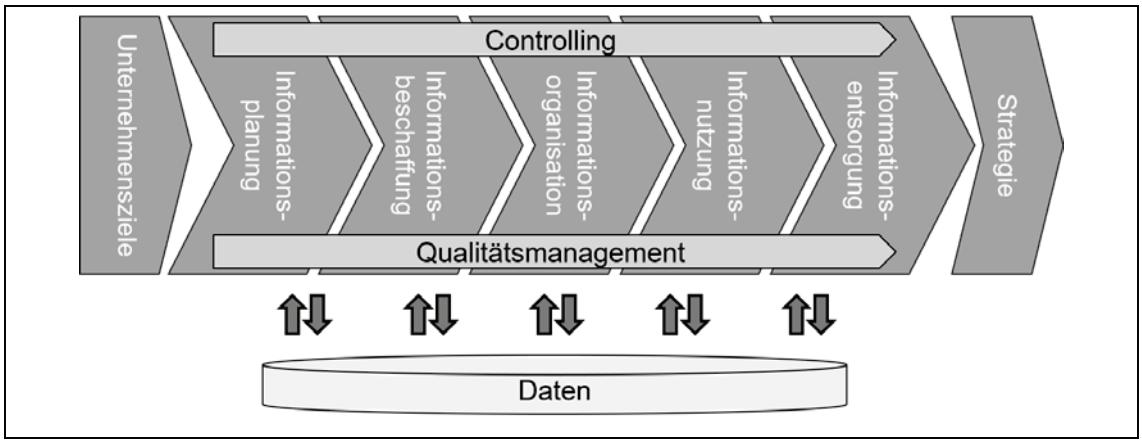

Abbildung 12: Informationsmanagementprozess[189]

Eine ähnliche Sichtweise hat auch KRCMAR, der das Informationsmanagement als Teil der Unternehmensführung sieht, welches die Ressource Information so einsetzen soll,

[185] Brust, O.-E. et al. (2015), S. 360.
[186] Christmann, A. (1988), S. 68. zitiert nach; Voß, S.; Gutenschwager, K. (2001), S. 58.
[187] Picot, A.; Reichwald, R. (1991), S. 264. zitiert nach; Voß, S.; Gutenschwager, K. (2001), S. 59.
[188] Dippold, R. et al. (2005), S. 258.
[189] Quelle: in Anlehnung an Dippold, R. et al. (2005), S. 15.

dass die Unternehmensziele möglichst effektiv erreicht werden. Es geht um die Planung, Steuerung und Kontrolle von Informationssystemen und deren Technologien. Als eine Subaufgabe dieser Planungs- und Steuerungsaufgaben ergibt sich die Koordination von Informationsströmen.[190] Wobei hier kritisch angemerkt wird, dass das Datenmanagement eine ähnliche Aufgabe hat. Die Datenarchitektur und Datenmodellierung beschäftigen sich ebenfalls mit Planungs-, Steuerungs-, und Koordinationsaufgaben und dem Austausch oder auch Fluss, der Daten.

Das Informationsmanagement wird aus Sicht der Arbeit vom Datenmanagement durch die Konkretisierung des Zwecks der Daten abgegrenzt. Das Datenmanagement konzipiert und koordiniert die Grundlagen für die generelle Sammlung, Verarbeitung und Verteilung von Daten. Das Informationsmanagement wählt basierend auf den Unternehmenszielen, welche Daten wie, wann und wo gebraucht und verwendet werden, um den Beitrag zum Unternehmenserfolg zu maximieren. Die Definition folgt somit dem effizienten und effektiven Einsatz der Ressource Information.

Ein wesentlicher Punkt des Informationsmanagements ist die Koordination der Informationsflüsse. Somit ist die Informationslogistik als ein Teilgebiet des Informationsmanagements zu sehen. Es können die logistischen Grundprinzipien angewandt werden.[191] Die richtigen Informationen (vom Empfänger verstanden und benötigt) zum richtigen Zeitpunkt (für die Entscheidungsfällung), notwendig in der richtigen Menge (so viel wie nötig, so wenig wie möglich), am richtigen Ort (beim Empfänger verfügbar) in der richtigen Qualität (ausreichend detailliert und wahr, unmittelbar verwendbar).[192] Diese Definition unterstreicht die hier verwendete Abgrenzung von Daten- und Informationsmanagement. Das Informationsmanagement muss die Leistung – hier die richtigen Informationen – zur richtigen Zeit, am richtigen Ort und in der richtigen Menge liefern.

In Industrie 4.0 hat der Fluss von Informationen oder die Ermöglichung des Flusses von Daten eine zentrale Bedeutung. Dieser Fluss sollte entlang der Wertschöpfungskette und durch die Hierarchieebenen gewährleistet werden, um einen größtmöglichen Nutzen aus den Daten ziehen zu können. Dieses Prinzip wird in der Industrie 4.0 generell als horizontale und vertikale Integration verstanden.

Die horizontale Integration beschreibt die Vernetzung zum Daten- und Informationsaustausch entlang der Prozessebene oder Wertschöpfungskette. Diese Vernetzung kann unternehmensintern und/oder zwischen Unternehmen erfolgen. Die vertikale Integration sichert den Austausch innerhalb eines Unternehmens über die Hierarchieebenen hinweg. Ob es sich dabei um den Austausch zwischen operativ, taktisch und strategisch oder MES-Ebene, ERP-Ebene und SCM-Ebene handelt ist dabei unerheblich. Eine Eigenschaft ist dabei, dass die Daten der untersten Ebene der Hierarchie nach aufsteigend stärker - zu Informationen - verdichtet werden.[193]

Abschließend wird zusammenfassend festgestellt, dass das Informationsmanagement sich mit der zielgerichteten Planung und Befriedigung des Informationsbedarfs

[190] Vgl. Krcmar, H. (2015), S. 1.
[191] Vgl. Krcmar, H. (2015), S. 117.
[192] Vgl. Augustin, S. (1990)
[193] Vgl. Lahrmann, G.; Stroh, F. (2008), S. 145.

beschäftigt. Diese Planung geht entlang der Wertschöpfungskette und über die Hierarchiestufen hinweg und gibt somit die Rahmenbedingungen für das Datenmanagement vor.

Nicht zu vernachlässigen ist die Rolle eines gut abgestimmten Daten- und Informationsmanagement für die Datenqualität. Es muss sichergestellt werden, dass die Daten und Informationen so übertragen und weiterverarbeitet werden, dass sie für die Endanwendung in ausreichender Qualität vorliegen. Im folgenden Abschnitt wird daher die Daten- und Informationsqualität genauer beschrieben.

3.3 Daten- und Informationsqualität

Die ISO definiert Qualität als „Gesamtheit von Eigenschaften und Merkmalen eines Produktes oder einer Dienstleistung, die sich auf deren Eignung zur Erfüllung festgelegter oder vorausgesetzter Erfordernisse bezieht“.[194]

Der Definition folgend kann eine Ausprägung der Qualität nur dann bestimmt werden, wenn vorher Kriterien bzw. deren Abstufungen festlegt werden mit denen diese gemessen werden soll. Wird das Kriterium zu einem gewissen Prozentsatz erfüllt, kann durch vorher definierte Grenzen ein Adjektiv wie *sehr gut*, *gut*, *befriedigend*, *mangelhaft* oder *ungenügend* der Qualität vorangesetzt werden. Bei der Datenqualität stellt sich das gleiche Problem wie bei der Qualität selbst: Es müssen Kriterien definiert werden, anhand derer die Datenqualität gemessen werden kann. In diesem Zusammenhang spricht man von *fit for use*. Die Daten, die verwendet werden, müssen eine hohe Qualität haben, die restlichen sind nicht relevant. In der Literatur gibt es drei Ansätze, zur Festlegung der Kriterien zur Messung der Datenqualität: der Intuitive, der Theoretische und der Empirische. Der intuitive Ansatz liegt dann vor, wenn auf Grund der Erfahrung bzw. des Wissens feststellt wird, welche Kriterien wichtig sind. Der theoretische Ansatz legt seinen Schwerpunkt darauf, wann die Daten in ihrer Erstellung mangelhaft werden. Der empirische Ansatz hält fest, welche Attribute für den Datennutzer bedeutend sind. Jeder Ansatz hat Vor- und Nachteile. Die Vorteile des intuitiven Ansatzes liegen in der Anpassungsfähigkeit. Je nach Ziel der Studie bzw. des Vorhabens können die Kriterien angepasst werden. Ebenfalls kann der theoretische Ansatz eine umfassende Reihe von Attributen liefern, die wesentlich für das Produkt Daten ist. Allerdings beziehen beide Ansätze nicht den Endnutzer dieser Daten ein, was dazu führt, dass theoretische Ansätze zur Verbesserung der Datenqualität verglichen mit empirischen Ansätzen schlechter abschneiden.[195] Mit dem *fit for use* - Ansatz wird Datenqualität wie folgt definiert werden:

Daten sind qualitativ, wenn sie den Anforderungen der vorgesehenen Verwendung entsprechen. Sie sind nicht qualitativ, wenn sie das nicht tun. Die Qualität hängt vom Verwendungszweck wie auch von den Daten selbst ab.[196]

[194] ÖNORM EN ISO 9000:2015-11-15 (2015)
[195] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 6 ff.
[196] Vgl. Olson, J. E. (2003), S. 24.

3.3.1 Datenqualitätsdarstellung

Aufgrund der *fit for use* Herangehensweise wurden laufend angepasste Datenqualitätsgrößen definiert. Diese waren untereinander nicht vergleichbar oder redundant. WANG und STRONG haben 1996 mit einer umfangreichen Literaturstudie und Umfrage ein Rahmenwerk veröffentlicht, welches die Datenqualität mit definierten Größen messbar machen soll. Es wurden vier Datenqualitätskategorien und 15 Datenqualitätsdimensionen definiert.[197]

Basierend auf diesen Ergebnissen wurde von der DGIQ[198] das Rahmenwerk adaptiert und ausformuliert. Dieses Rahmenwerk fand Eingang in viele deutschsprachige Bücher über Daten- und Informationsqualität. Bei den folgenden Definitionen ist zu beachten, dass Daten und Informationen in den Ausführungen der DGIQ synonym verwendet wurden.[199] Tabelle 10 zeigt die 15 Datenqualitätsdimensionen gruppiert in die vier Datenqualitätskategorien.

Die synonyme Verwendung von Daten und Informationen wird in diesem Abschnitt übernommen. In weiterer Folge wird ausschließlich von der Datenqualität geschrieben, da Daten das Vorprodukt von Informationen sind. Somit können Informationen nur von hoher Qualität sein, wenn die Daten von hoher Qualität sind. Eine strenge und eigentlich künstliche und nicht notwendige Abgrenzung der Begriffe wird als unnötig erachtet.

Tabelle 10: Datenqualitätskategorien und -dimensionen[200]

Datenqualität				
Kategorien	**Inhärent**	**Zweckabhängig**	**Darstellungs-bezogen**	**Systemgestützt**
Dimensionen	Glaubwürdigkeit	Wertschöpfung	Eindeutige Auslegbarkeit	Zugänglichkeit
	Fehlerfreiheit	Relevanz	Verständlichkeit	Bearbeitbarkeit
	Objektivität	Aktualität	Einheitliche Darstellung	
	Hohes Ansehen	Vollständigkeit	Übersichtlichkeit	
		Angemessener Umfang		

Datenqualitätskategorien

Die vier Kategorien fassen ähnliche Datenqualitätsdimensionen zusammen und bilden einen Untersuchungsgegenstand ab, den Daten zugeschrieben werden.[201]

[197] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 19 f.
[198] DGIQ: Deutsche Gesellschaft für Informationsqualität
[199] Vgl. Rohweder, J. P. et al. (2015), S. 26.
[200] Quelle: in Anlehnung an Rohweder, J. P. et al. (2015), S. 28 ff. zitiert nach Bernerstätter, R.; Kühnast, R. (2018), S. 3.
[201] Vgl. Rohweder, J. P. et al. (2015), S. 30 f.

Bei der systemunterstützenden Datenqualitätskategorie werden die Systeme zur Datenverarbeitung und -bearbeitung untersucht. Die inhärente Datenqualitätskategorie befasst sich mit dem Inhalt der Daten an sich. Die darstellungsbezogene Kategorie kann unter der Berücksichtigung der Darstellung der Daten beurteilt werden. Der Aspekt der eigentlichen Nutzung der Daten wird mit der zweckabhängigen Kategorie beschrieben.[202]

Abbildung 13 zeigt das Zusammenwirken der Datenqualitätskategorien grafisch. Hier ist die eigentliche Schwäche der derzeitig vorherrschenden Definition und Aufarbeitung der Datenqualitätskategorien und infolge dessen deren zugeordneten Datenqualitätsdimensionen zu sehen. Die Auseinandersetzung erfolgte im Fokus von Business Intelligence Anwendungen. Industriedaten des Produktionsprozesses sind nicht bewertet worden.

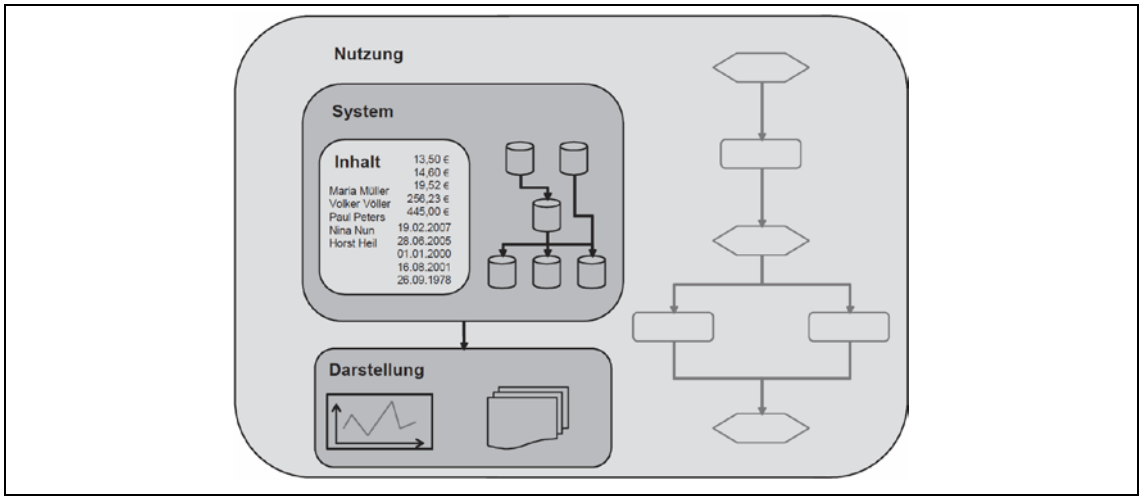

Abbildung 13: Untersuchungsgegenstände der Datenqualität[203]

Datenqualitätsdimensionen

Eine Datenqualitätsdimension fasst eine Menge von Datenqualitätseigenschaften zusammen, die einen Aspekt der Datenqualität darstellen.

Zugänglichkeit (accessibility):

„Informationen sind zugänglich, wenn sie anhand einfacher Verfahren und auf direktem Weg für den Anwender abrufbar sind.“[204]

Da Daten früher in Papierform abgelegt wurden, ist die Zugänglichkeit bei der Erhebung der Datenqualitätsdimensionen durch WANG und STRONG als eine relevante identifiziert worden. Die Übernahme in ein Rahmenwerk in der heutigen hochdigitalisierten Zeit ist jedoch gerechtfertigt, da durch Systembrüche und Zugangsrechte, die Zugänglichkeit von Daten nach wie vor relevant ist.[205]

[202] Vgl. Rohweder, J. P. et al. (2015), S. 30 f.
[203] Quelle: Rohweder, J. P. et al. (2015), S. 32.
[204] Rohweder, J. P. et al. (2015), S. 32.
[205] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 21.

Entlang der horizontalen und vertikalen Datenintegration sollten keine Barrieren bestehen, da diese den Datenaufnahmeprozess für Analyseprozesse erschweren und die spätere automatische Implementierung erschweren oder unmöglich machen können.

Angemessener Umfang (appropriate amount of data):

„Informationen sind von angemessenem Umfang, wenn die Menge der verfügbaren Information den gestellten Anforderungen genügt.“[206]

Die Anzahl von Datensätzen und Attributen soll für den Anwendungsfall ausreichend sein. In der ursprünglichen Definition der Dimension ist auch ein Mehr an Daten als schlechter Einfluss auf die positive Ausprägung zu verstehen.[207]

Diese Einschränkung muss unter BI-Aspekten verstanden werden, da eine Darstellung von zu vielen Daten kontraproduktiv für das Verständnis ist. In der Rückmeldung von Informationen sind zu viele Daten auch kontraproduktiv. Werden Anlagenausfälle mit Stör- und Schadenscodes zurückgemeldet, so führt eine zu große Auswahl an Standardcodes dazu, dass falsche genommen werden. Das wirkt sich auf die Datenqualität aus.[208]

Im industriellen Umfeld und bei Big Data Analytik Anwendungen sollten möglichst viele Daten vorliegen, um Algorithmen zu ermöglichen Muster zu erkennen. Der Fokus liegt nicht auf der Darstellung der Daten. Trotzdem muss im Rahmen des Informationsmanagements und der Informationslogistik dafür gesorgt werden, dass ausreichend Daten für die Analysezwecke, die hier die Endanwender sind, vorliegen.

Glaubwürdigkeit (believability):

„Informationen sind glaubwürdig, wenn Zertifikate einen hohen Qualitätsstandard ausweisen oder die Informationsgewinnung und verbreitung [sic!] mit hohem Aufwand betrieben werden.“[209]

Im ursprünglichen Sinn geht es um ein gutes Marketing für Informationen. So sind Daten aus den staatlichen Statistikämtern glaubwürdiger als Daten aus einem Boulevardmedium. Es werden somit die Vertrauenswürdigkeit und Zuverlässigkeit der Datenlieferanten bewertet.[210] In der Instandhaltung sind Daten, die durch definierte Prozesse und automatisierte System aufgenommen werden, glaubwürdiger als wenn diese Voraussetzungen nicht erfüllt sind.[211]

In Big Data Anwendungen des industriellen Umfeldes spielt die Glaubwürdigkeit in einer ersten Betrachtung eine untergeordnete Rolle. Es darf jedoch nicht vernachlässigt werden, dass Analysen mit erheblichem Zeitaufwand verbunden sind. Glaubwürdige Datenquellen erhöhen die Wahrscheinlichkeit, dass das erhaltene Ergebnis repräsentativ für die Situation ist.

[206] Rohweder, J. P. et al. (2015), S. 33.
[207] Vgl. Rohweder, J. P. et al. (2015), S. 33.
[208] Vgl. Woodall, P. et al. (2015), S. 331 f.
[209] Rohweder, J. P. et al. (2015), S. 34.
[210] Vgl. Rohweder, J. P. et al. (2015), S. 34.
[211] Vgl. Woodall, P. et al. (2015), S. 332.

Vollständigkeit (completeness):

„Informationen sind vollständig, wenn sie nicht fehlen und zu den festgelegten Zeitpunkten in den jeweiligen Prozess-Schritten zur Verfügung stehen.“[212]

Durch die Definition erübrigt sich eine eigene Dimension zur Pünktlichkeit, die als Zeitnähe bezeichnet wird.[213] Die Vollständigkeit wird in unterschiedlichen Ausprägungen gemessen. Wichtig ist eine Definition einer Vergleichsmenge, gegen die der Vergleich auf Vollständigkeit durchgeführt wird,[214] da eine prinzipielle, quasi axiomatische, Feststellung von Vollständigkeit nicht möglich ist.[215] EMRAN definiert die Vollständigkeit als multidimensionales Konzept, welches es erfordert eine Messgrundlage zu definieren, die das Fehlende feststellt. Er führte dafür eine ausführliche Literaturstudie durch, wie Vollständigkeit in Bezug auf Daten und Informationen definiert wird.[216] HEINRICH und KLIER schränken die Vollständigkeit auf das Fehlen von Einträgen ein und bestimmen damit die Qualität der Daten in dieser Dimension.[217]

Vollständige Daten sind auch im industriellen Umfeld wesentlich, um Analysen durchzuführen. Es ist weniger die Pünktlichkeit ausschlaggebend, als dass alle Datenfelder einer Aufzeichnung ausgefüllt sind. Die durchgängige Datenmodellierung und automatische Datenerfassungssysteme unterstützten hier wesentlich.

Übersichtlichkeit (concise representation):

„Informationen sind übersichtlich, wenn genau die benötigten Informationen in einem passenden und leicht fassbaren Format dargestellt sind.“[218]

Hier wird eine Dimension gemessen, die speziell für den Endanwender von bereitgestellten Informationen relevant ist. Daten und Informationen sind gut dargestellt, wenn sie übersichtlich, also nicht in überwältigender Form angezeigt werden.[219] Es handelt sich hier um eine spezielle Datenqualitätsdimension, die für die menschliche Interpretation und Verarbeitung wichtig ist. Im industriellen Umfeld sollte diese Dimension jedoch keine Rolle spielen, da ohnehin von Big Data ausgegangen wird und die Zusammenhänge so komplex sind, dass automatische Algorithmen verwendet werden, um Zusammenhänge und Muster zu erkennen.

Einheitliche Darstellung (consistent represantation):

„Informationen sind einheitlich dargestellt, wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden.“[220]

Daten und Informationen sind einheitlich dargestellt, wenn sie unabhängig vom Zeitpunkt und von der Art der Datenerfassung im gleichen Format, Layout und mit demselben Wertevorrat dargestellt und erfasst werden. Der Datenerfasser gilt als Schwachpunkt,

[212] Rohweder, J. P. et al. (2015), S. 34.
[213] Vgl. Apel, D. et al. (2015), S. 25.
[214] Vgl. Rohweder, J. P. et al. (2015), S. 34.
[215] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 7.
[216] Vgl. Emran, N. A. (2015), S. 120 f.
[217] Vgl. Heinrich, B.; Klier, M. (2015), S. 54 f.
[218] Rohweder, J. P. et al. (2015), S. 35.
[219] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 21.
[220] Rohweder, J. P. et al. (2015), S. 36.

da dieser nach Ort, Kulturkreis und Tagesverfassung Daten unterschiedlich aufzeichnen kann.[221] Daten können fehlerfrei und eindeutig auslegbar sein und trotzdem nicht die Dimension der einheitlichen Darstellung erfüllen. Die Dimension spielt eine wesentliche Rolle bei der Datenaggregation.[222]

Die einheitliche Darstellung ist im industriellen Umfeld von Big Data nicht zu unterschätzen. Die Variety von Big Data führt zu Daten aus unterschiedlichen Quellen. Deren Inhalt muss einheitlich abgebildet werden, damit ein einheitlicher Datenbestand für die Analysen erzeugt wird.

Bearbeitbarkeit (ease of manipulation):

„Informationen sind leicht bearbeitbar, wenn sie leicht zu ändern und für unterschiedliche Zwecke zu verwenden sind.“[223]

Eine gute Bearbeitbarkeit ist positiv und negativ zugleich. Positiv ist die leichte Anpassung und folglich die universelle Verwendbarkeit der Daten zu betrachten. Negativ ist die Gefahr einer leichten Verfälschung der Daten. Berechtigte Nutzer müssen die Daten leicht verändern und unberechtigte Nutzer dürfen die Daten nicht verändern können. Dann ist die Bearbeitbarkeit als gut einzustufen.[224]

Für die Anwendungen im Big Data Umfeld bei industriellen Anwendungen wird die Bearbeitbarkeit nicht negativ gesehen. Daten müssen im vorliegenden Format leicht veränderbar sein, um nötige Transformationsschritte durchführen zu können. Der negative Aspekt der möglichen Veränderung durch Mitarbeiter wird in der Glaubwürdigkeit gesehen und sollte nicht mit der Bearbeitbarkeit bewertet werden.

Fehlerfreiheit (accuracy):

„Informationen sind fehlerfrei, wenn sie mit der Realität übereinstimmen.“[225]

Abbildung 14 gibt einen Überblick, was fehlerfreie Daten bedeutet. Von allen möglichen validen Werten (z. B. korrekter Wertebereich), sind jene Daten fehlerfrei die richtige Werte haben (korrekte Messergebnisse) und richtig dargestellt sind.

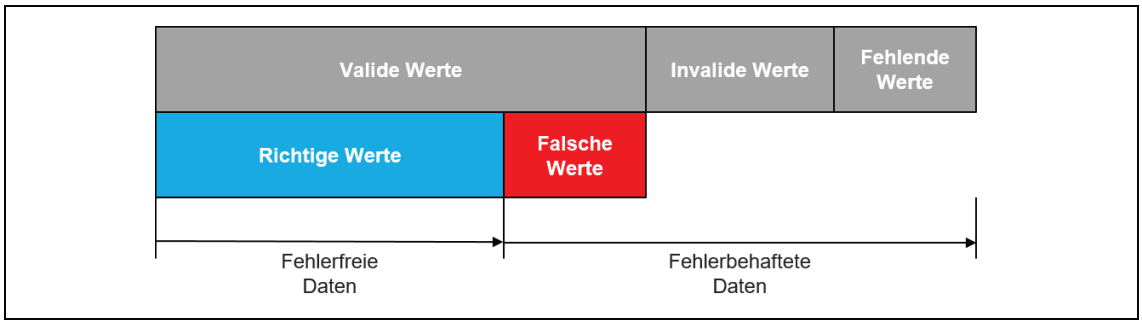

Abbildung 14: Fehlerfreie Daten heruntergebrochen[226]

[221] Vgl. Rohweder, J. P. et al. (2015), S. 36.
[222] Vgl. Olson, J. E. (2003), S. 29.
[223] Rohweder, J. P. et al. (2015), S. 37.
[224] Vgl. Rohweder, J. P. et al. (2015), S. 37 f.
[225] Rohweder, J. P. et al. (2015), S. 37.
[226] Quelle: in Anlehnung an Olson, J. E. (2003), S. 33.

Gespeicherte Daten bilden die Realität vereinfacht, gröber oder in aggregierter Form ab. Die Abbildung darf jedoch nicht der Realität widersprechen, weshalb die Übereinstimmung mit der Realität als Qualitätsmerkmal verwendet wird.[227] Die Fehlerfreiheit als Genauigkeit gehört zu den meistverwendeten Datenqualitätsdimensionen.[228]

Eine gut definierte Datenarchitektur und -modellierung ist ein Schlüsselfaktor, um invalide Werte und fehlende Werte zu vermeiden. Diese können nicht falsch ausgefüllt werden, wenn die Vorgaben es nicht erlauben, was bereits einen großen Teil möglicher Ungenauigkeit vermeidet.

Eindeutige Auslegbarkeit (interpretability):

„Informationen sind eindeutig auslegbar, wenn sie in gleicher, fachlich korrekter Art und Weise begriffen werden."[229]

Die Daten müssen in der Verarbeitungssupplychain, von Datenerzeuger bis zum Datennutzer, einheitlich interpretiert werden. Dazu müssen sie klar formuliert und mit geeigneten Symbolen und verständlicher Sprache abgebildet und weitergegeben werden.[230]

Ein Problem kann darin liegen, dass Annahmen bei Berechnungen unterschiedlich sind, da diese nicht eindeutig definiert sind. Beispielsweise die Durchlaufzeit, die unterschiedlichen Bezugspunkte zur Berechnung haben kann. Wenn mit der Wiederbeschaffungszeit von Ersatzteilen mit unterschiedlichen Definitionen gearbeitet wird, kann das zu Anlagenstillständen führen, da die Ersatzteile nicht mehr verfügbar sind.[231]

Objektivität (objectivity):

„Informationen sind objektiv, wenn sie streng sachlich und wertfrei sind."[232]

Rohdaten sind als objektiv zu bewerten, da sie frei von subjektiven Interpretationen und Arbeitsschritten sind. Um die Objektivität zu bewerten, muss der verarbeitende Prozess bekannt sein. Solche Prozesse können Aggregationen oder Transformationen der Daten sein.[233]

Metadaten unterstützen bei der Beurteilung der Objektivität der Daten, wenn es um die Bewertung der Aggregations- und Transformationsschritte geht. Die Objektivität der Datenerfassung wird bereits in der Glaubwürdigkeit bewertet. Es wird daher bezweifelt, dass diese Datenqualitätsdimension im industriellen Umfeld relevant ist, da die wesentliche Frage mit der Glaubwürdigkeit abgedeckt ist.

[227] Vgl. Rohweder, J. P. et al. (2015), S. 37.
[228] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 7.
[229] Rohweder, J. P. et al. (2015), S. 38.
[230] Vgl. Rohweder, J. P. et al. (2015), S. 38.
[231] Vgl. Woodall, P. et al. (2015), S. 330.
[232] Rohweder, J. P. et al. (2015), S. 39.
[233] Vgl. Rohweder, J. P. et al. (2015), S. 39.

Relevanz (relevancy):

„Informationen sind relevant, wenn sie für den Anwender notwendige Informationen liefern."[234]

Die Notwendigkeit einer Information definiert sich durch die erhöhte Wahrscheinlichkeit der Zielerreichung, wenn die Information verwendet wird. Ist dieser Fall gegeben, so ist die Information notwendig und folglich relevant.[235] Die Relevanz kann auch als ein Ausmaß gesehen werden, wie sehr die Daten für den Nutzer verwendbar und hilfreich sind.[236]

Wird die Datenmenge betrachtet, die bei Big Data Anwendungen anfallen, erhält die Relevanz der Daten ein großes Gewicht. Viele Daten können die Rechenzeit bei Analysen stark erhöhen, was speziell bei Real-Time Auswertungen problematisch werden kann. Es muss daher bewertet werden, welche Daten relevant sind. Das ist eine wichtige Aufgabe in der Zieldefinition des Informationsmanagements.

Hohes Ansehen (reputation):

„Informationen sind hoch angesehen, wenn die Informationsquelle, das Transportmedium und das verarbeitende System im Ruf einer hohen Vertrauenswürdigkeit und Kompetenz stehen."[237]

Das hohe Ansehen von Quellen und Übertragungsmedien muss über einen längeren Zeitraum erworben werden. Zeigt die Erfahrung, dass diese zuverlässig sind, dann steigt damit die Reputation. Hohes Ansehen ist als Datenqualitätsdimension dann wichtig, wenn andere beeinflusste Dimensionen, wie z. B. die Fehlerfreiheit, nicht gemessen werden können.[238]

Die Grenze zwischen hohem Ansehen und Glaubwürdigkeit ist schwer zu ziehen. Im Sinne der eindeutigen Bewertung und Handhabbarkeit eines Modells, das die Datenqualitätsdimensionen umfasst, werden diese beiden Dimensionen zusammengefasst und es wird die Glaubwürdigkeit als relevante Datenqualitätsdimension verwendet.

Aktualität (timeliness):

„Informationen sind aktuell, wenn sie die tatsächliche Eigenschaft des beschriebenen Objektes zeitnah abbilden."[239]

Die Aktualität grenzt sich von der Pünktlichkeit - welche ein Teil der Dimension Vollständigkeit ist - ab indem es dabei um die Anpassung an die veränderten Rahmenbedingungen der realen Welt geht. Ein stärkerer Konnex besteht hingegen zur Dimension Fehlerfreiheit.[240] Daten sind auch aktuell, wenn sie einen möglichst direkten

[234] Rohweder, J. P. et al. (2015), S. 39.
[235] Vgl. Rohweder, J. P. et al. (2015), S. 39.
[236] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 31.
[237] Rohweder, J. P. et al. (2015), S. 40.
[238] Vgl. Rohweder, J. P. et al. (2015), S. 40.
[239] Rohweder, J. P. et al. (2015), S. 41.
[240] Vgl. Rohweder, J. P. et al. (2015), S. 41.

Bezug zur Gegenwart haben. Es wird somit die Frage beantwortet inwieweit der Datenbestand die Gegebenheiten der Realität abbildet.[241]

Hier spiegelt sich eindeutig der Ursprung der Datenqualitätsdimensionen wider. Der Fokus lag vor allem auf Stammdaten, die über die Zeit an Aktualität verlieren können. Für diese Arbeit wird die Aktualität der Daten auf den Rückmeldezeitpunkt bezogen. Es wird bewertet wie nahe der Moment der Datenerzeugung, der Moment der Datenaufzeichnung und der Moment der Datenspeicherung auseinanderliegen. Die Aktualität dient somit als ein Indikator für die Zeitstempelzuverlässigkeit.

Verständlichkeit (understandability):

„Informationen sind verständlich, wenn sie unmittelbar von den Anwendern verstanden und für deren Zwecke eingesetzt werden können."[242]

Damit Daten als gut dargestellt gelten, müssen sie neben einer einfachen Interpretierbarkeit, auch einfach verständlich sein. Die Daten sollten daher nicht zweideutig sein und leicht aufgenommen werden können.[243]

Die Verständlichkeit ist für die industrielle Anwendung bei Big Data Analysen nicht von Relevanz. Alle nötigen Aspekte werden durch die *einheitliche Darstellung* und die *eindeutige Auslegbarkeit* abgebildet. Beide Dimensionen sorgen dafür, dass Algorithmen Daten korrekt für die Mustererkennung verwenden.

Wertschöpfung (value-added):

„Informationen sind wertschöpfend, wenn ihre Nutzung zu einer quantifizierbaren Steigerung einer monetären Zielfunktion führen kann."[244]

Liefern Daten und Informationen einen Beitrag zu einer Entscheidung, die monetär eine positive Auswirkung hat, so sind die Daten und Informationen wertschöpfend. Der konkrete Wertschöpfungsbeitrag, ist die Differenz einer monetären Zielfunktion, ohne der Entscheidung und mit der Entscheidung die mit Hilfe der Daten getroffen wurde.[245] Die Abgrenzung zur Relevanz liegt darin, dass relevante Daten nicht zwingend einen messbaren Wertschöpfungsbeitrag liefern.[246]

3.3.2 Datenqualitätsmessung

Zu den Dimensionen können je nach Einsatzzweck Metriken definiert werden, um die Datenqualität bezogen auf die Dimension zu messen. Mithilfe dieser Metriken ist es möglich Verbesserungen, die sich im Rahmen des Datenqualitätsmanagements ergeben quantitativ zu erheben und diese in weiterer Folge monetär zu bewerten. KLIER hat neben verschiedener Metriken auch Anforderungen definiert denen Datenqualitätsmetriken entsprechen müssen. Datenqualitätsmetriken müssen demnach normiert sein, um die Interpretierbarkeit und Vergleichbarkeit untereinander zu

[241] Vgl. Heinrich, B.; Klier, M. (2015), S. 62.
[242] Rohweder, J. P. et al. (2015), S. 42.
[243] Vgl. Wang, R. Y.; Strong, D. M. (1996), S. 21 & 32.
[244] Rohweder, J. P. et al. (2015), S. 42.
[245] Vgl. Rohweder, J. P. et al. (2015), S. 42.
[246] Vgl. Rohweder, J. P. et al. (2015), S. 40.

gewährleisten. Des Weiteren ist eine kardinale Skalierung vorausgesetzt, damit die zeitliche Entwicklung betrachtet werden kann und dem Ergebnis der Messung mit der Metrik ein ökonomischer Wert zugeordnet werden kann. Die dritte Voraussetzung ist die Sensibilisierbarkeit, die sich auf die oben erwähnte Anpassung an die Anforderungen bezieht. Die Aggregierbarkeit soll sicherstellen, dass die Metrik auf Datensatzebene, wie auch auf Datenbankebene anwendbar ist. Um die Metriken praktisch anwenden zu können, ist die Operationalisierbarkeit mittels Messverfahren eine weitere Anforderung an Datenqualitätsmetriken. Die letzte Anforderung ist die fachliche Interpretierbarkeit, da die reine Anwendung der Normierung und der Kardinalität meist nicht ausreicht und die Ergebnisse von einem Experten interpretiert werden müssen.[247]

Auch wenn die Datenqualitätsdimensionen durch das Rahmenwerk von WANG und STRONG und jenes der DGIQ vereinheitlicht sind, werden die Metriken für die Messung der Datenqualität in den einzelnen Dimensionen oft situationsbedingt erstellt.

Mögliche generische Metriken sind folgende:

- Anzahl der Zeilen die zumindest einen falschen Wert aufweisen.
- Anzahl der Verletzungen der Schlüsselintegrität (nicht redundante Primärschlüssel, Fremdschlüssel ohne Primärschlüssel).
- Auflistung der Fehler welche durch den falschen Speicherort entstehen.
- Auflistung der Fehler in Bezug auf das Erzeugungszeitpunkt der Daten.
- Anzahl der Verletzungen von Datenregeln.
- Anzahl der Fehler je Datenelement.

Abgerufene oder gelieferte Daten können mit Metriken objektiv bewertet werden. Neben diesen Vorteilen gibt es eine Reihe von Nachteilen von Datenqualitätsmetriken. Sie sind in den wenigsten Fällen exakt. Schlecht durchdachte Metriken zeigen die Datenqualität übermäßig schlecht oder gut an. Sie identifizieren die Ursachen von Probleme nicht, sondern zeigen deren Existenz auf. Der Schluss aus einer implementierten Metrik darf somit nicht sein, dass die Arbeit der Datenqualitätsverbesserung getan ist, da keine Verbesserungsmaßnahmen von der Metrik allein abgeleitet werden können.[248]

Viele Metriken der Literatur sind unzureichend definiert und führen zu zweideutigen Ergebnissen. So prüft die Metrik Korrektheit die Qualität der Dimension Fehlerfreiheit. Werden Namen verglichen, dann wird der Eintrag „Meierhofre“ mit dem korrekten Namen „Mayerhofer“ verglichen. Der Vergleich stellt fest, auf wie vielen Stellen es Unterschiede gibt; in diesem Fall vier. Vergleicht man den Eintrag „Wein“ mit dem Namen „Mayr“, würde man das gleiche Ergebnis für die Korrektheit erhalten.[249]

Dieses plakative Beispiel zu Schwächen von Datenqualitätsmetriken zeigt wie schwierig es ist aussagekräftige Metriken zu finden. Ein weiteres Problem ist die Definition des Vergleichswertes. Dieser muss im beschriebenen Fall für jeden Namen in einer Namensdatenbank gegeben sein. Der administrative Aufwand wäre enorm und es könnte nur eine Stichprobenprüfung durchgeführt werden.

Eine Metrik zur Dimension Vollständigkeit ist im Gegensatz einfacher zur formulieren. Es wird gezählt wie oft ein Wert in einem Attribut fehlt und ggf. in Verhältnis mit der

[247] Vgl. Klier, M. (2008), S. 225 f.
[248] Vgl. Olson, J. E. (2003), S. 82 ff.
[249] Vgl. Klier, M. (2008), S. 227.

Gesamtanzahl an Werten gesetzt. Der Hinweis auf fehlende Werte wird in der Regel mit *NULL* gegeben. Es muss beachtet werden, dass die Möglichkeit besteht, dass der Wert nicht fehlt, sondern nicht existiert. Dafür muss eine Notation vereinbart werden, die diesen Fall anzeigt. Falls nicht nur einzelne Attribute bewertet werden sollen, sondern ganze Datenbestände, ist es möglich die Attribute zu gewichten, da es vorkommen kann, dass die Vollständigkeit eines Attributes wichtiger ist als die eines anderen.[250] Auch dieses Beispiel zeigt, dass die Definition von Metriken in Spezialfällen nicht trivial ist und viel Kenntnis des Datenbestandes erfordert.

Eine generische Definition einer Metrik für Vollständigkeit nimmt EMRAN vor indem er ein betrachtetes Datenset ins Verhältnis zu einem Referenzdatenset setzt, wie in Formel 3-1 zu sehen ist. Abhängig vom Anwendungsfall muss bestimmt werden, wie fehlende Daten definiert werden und was als Referenzdatenset verwendet wird.[251]

D…Betrachtetes Datenset
R…Referenzdatenset

$$Vollständigkeit(D,R) = \frac{|D \cap R|}{|R|} \qquad 3\text{-}1$$

Wird versucht, diese Metrik auf Sensormessungen oder Prozessdaten umzulegen, stellt sich die Frage, welcher Vergleichswert verwendet werden soll. Die Datenqualität ist in diesen Fällen mit anderen Methoden festzustellen.

3.3.3 Datenqualitätsmanagement

Ähnlich wie bei klassischen Produktionsfaktoren und -prozessen sollten auch Daten einem rigorosen Qualitätsmanagementprozess unterworfen werden. Dazu wurde der Begriff des *Total Data Quality Management (TDQM)* eingeführt. Es lehnt sich am Demingzyklus an und besteht aus den Schritten *Definieren, Messen, Analysieren und Verbessern*.[252] Ein Nachteil dieses einfachen Zyklus ist das Fehlen eines Control-Schrittes wie im DMAIC Zyklus aus Six Sigma.[253]

Mit der Messung der Datenqualität in den Datenqualitätsdimensionen kann ein gezieltes Datenqualitätsmanagement durchgeführt werden. Abbildung 15 zeigt mit dem Datenqualitätsregelkreis eine mögliche Umsetzung des TDQM. An dessen Beginn stehen die oben beschriebenen Datenqualitätsdimensionen und Datenqualitätsmetriken.

Die Datenqualitätsmaßnahmen führen zu einer Verbesserung der Datenqualität, gemessen an den Datenqualitätsmetriken. Dieser Verbesserung muss ein Nutzen zugeordnet werden, der durch die Metriken abgeschätzt werden kann. Da die Maßnahmen mit Kosten verbunden sind, sind jene Maßnahmen zu wählen, bei denen das Kosten-Nutzen Verhältnis am besten ist.[254]

[250] Vgl. Heinrich, B.; Klier, M. (2009), S. 36.
[251] Vgl. Emran, N. A. (2015), S. 122.
[252] Vgl. Wang, R. Y. (1998), S. 59 f.
[253] Vgl. Hazen, B. T. et al. (2014), S. 73.
[254] Vgl. Heinrich, B.; Klier, M. (2015), S. 50.

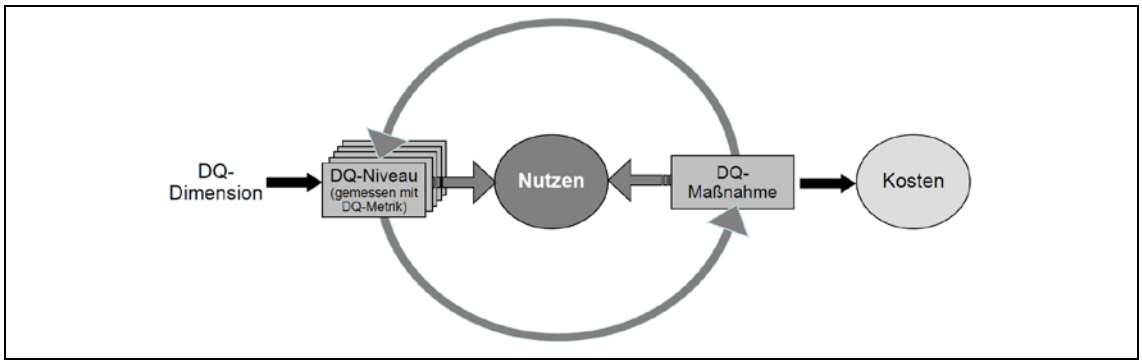

Abbildung 15: Datenqualitätsregelkreis[255]

Diese Herangehensweise hat einen stark operativen und kurzzyklischen Charakter. Es ist auch wichtig, dass das Datenqualitätsmanagement auf strategischer Ebene angesiedelt wird und langfristig ausgerichtet ist.[256] Dazu wurde ein Corporate Data Quality Modell entwickelt, welches relevante Gestaltungsobjekte realisiert. In den Ebenen Strategie, Organisation und Informationssystemarchitektur werden Verfahren definiert, die die Datenqualität nachhaltig verbessern und auf einem hohen Niveau halten sollen. Ein mögliches Verfahren ist der oben beschriebene Datenqualitätsregelkreis.[257]

Generell unterscheidet das Datenqualitätsmanagement zwischen reaktiven und präventiven Maßnahmen. Präventive Maßnahmen versuchen negative Auswirkungen der Datenqualität zu vermeiden, bevor diese eintreten. Die reaktiven Maßnahmen suchen nach Datenqualitätsproblemen. Der reaktive Ansatz hat neben anderen Nachteilen, jenen, dass Organisationen, die so vorgehen meist keine Metriken definiert haben, um die Datenqualität zu messen. Daher ist es kurzfristig schwierig Fehler zu finden. Die präventive Herangehensweise ist daher zu bevorzugen, da auch nachgewiesen wurde, dass diese kostengünstiger ist, als die reaktive. Eine einfache präventive Maßnahme sind Prüfregeln bei der Dateneingabe.[258] Diese sollten bei der Datenmodellierung im Rahmen des Datenmanagements bereits definiert werden.

Da Daten in allen Bereichen eines Unternehmens erzeugt werden ist für das Datenqualitätsmanagement die Besetzung von unterschiedlichen Rollen und Positionen nötig. Der Sponsor verantwortet den Aufbau einer Datenqualitätsstrategie. Es weiß um die Wichtigkeit von Datenqualität und deren Einfluss auf die Geschäftsprozesse und definiert Kennzahlen, um die Datenqualität zu messen. Er muss über die nötige Weisungsbefugnis über Mitarbeiter in der Datenqualitätsinitiative, über Ressourcen und finanzielle Mittel verfügen. In vielen Fällen obliegt diese Rolle den CIO (Chief Information Officer). Der Datenqualitätsbeauftragte oder Data Quality Officer, verantwortet die laufende Überwachung der Datenqualität. Er ist für die kontinuierliche Verbesserung verantwortlich. Der Datenverwalter sitzt in der jeweiligen Fachabteilung und besitzt viel Erfahrung in den Abläufen in der Abteilung und im Unternehmen. Seine Aufgaben sind operativer Natur.[259]

[255] Quelle: Klier, M. (2008), S. 224.
[256] Vgl. Otto, B.; Hinderer, H. (2009), S. 24.
[257] Vgl. Baumöl, U.; Meschke, M. (2009), S. 64.
[258] Vgl. Otto, B.; Österle, H. (2016), S. 32 f.
[259] Vgl. Weigel, N. (2015), S. 73 f.

Das Datenqualitätsmanagement sollte die logische Konsequenz sein, da Daten in der digitalisierten Wirtschaft des 21. Jahrhunderts immer bedeutender werden. Es bildet in gewisser Weise einen Rahmen für die vorliegende Arbeit, es ist jedoch nicht der Fokus. Das Reifegradmodell sollte dennoch dazu dienen, die Datenqualität laufend zu verbessern indem eine neue Reifegradstufe angestrebt und gehalten wird.

3.4 Kritische Würdigung der Datenqualität und Relevanz für die Arbeit

Die Datenqualität ist ein ausschlaggebender Faktor in der Analyse von Daten. So sehen 62% der Befragten einer Studie zu Industrial Data Analytics die Fehlerfreiheit von Daten als große bzw. sehr große Herausforderung in datenanalytischen Projekten.[260] Die bereits existierenden Datenqualitätsdimensionen und Vorgehensweisen des Datenqualitätsmanagements für Stamm- und Bewegungsdaten aus dem klassischen BI-Umfeld wurden noch nicht auf den industriellen Bereich übertragen. Eine unreflektierte Übertragung wird nicht möglich sein, da einige Datenqualitätsdimensionen explizit unter der Annahme des Kontaktes des Anwenders mit den Daten definiert wurden. Trotzdem zeigt sich bei genauerer Analyse der Formulierungen und Beschreibungen aus der Literatur, dass eine Definition von ausgewählten Datenqualitätsdimension für das industrielle Umfeld der Big Data Analytik möglich ist.

Im vorliegenden Kapitel wurden die Datenqualitätsdimensionen beschrieben und bereits auf die Anwendbarkeit im industriellen Umfeld überprüft und teilweise neu interpretiert. Die Erkenntnisse daraus werden hier zusammengefasst. Die systemgestützten Datenqualitätsdimensionen „*Zugänglichkeit*“ und „*Bearbeitbarkeit*“ haben auch im industriellen Umfeld der Datenanalyse ihre Berechtigung. Die Zugänglichkeit muss folgend präzisiert werden:

„Daten sind zugänglich, wenn sie durch digitale Informationssysteme von einer Anwendung automatisch abgerufen werden können.“

Der Fokus auf digitale Informationssysteme, bzw. rechnergestützte Systeme ist für industrielle Big Data Anwendungen wichtig. Die Zugänglichkeit hat für den späteren Einsatz von datenanalytischen Ergebnissen großen Wert, daher müssen die Daten automatisch durch eine Anwendung abgreifbar sein.

Die Bearbeitbarkeit muss nicht neu definiert werden. Die Feststellung, dass eine leichte Bearbeitbarkeit nicht negativ ist, ist ausreichend.

Die darstellungsbezogenen Datenqualitätsdimensionen haben für industrielle automatisierte Analysen von großen Datenmengen eine eingeschränkte Rolle. Wie bereits erörtert, werden die Dimensionen *Verständlichkeit* und *Übersichtlichkeit* nicht weiter benötigt. Die *Übersichtlichkeit* ist obsolet, da der Zweck für industrielle Anwendungen bereits durch die „*eindeutige Auslegbarkeit*“ und die „*einheitliche Darstellung*“ beschrieben ist. Diese beiden Datenqualitätsdimensionen müssen nicht weiter definiert werden.

[260] Vgl. Lueth, K. L. et al. (2016), S. 49.

Zweckabhängige Datenqualitätsdimensionen müssen ebenfalls fallweise angepasst werden. Die Relevanz ist konkreter für den Einsatz von Algorithmen und mathematischen Verfahren zu spezifizieren:

„Daten sind relevant, wenn sie für den Einsatz ausreichend Informationsgehalt haben."

Die Messung des Informationsgehaltes kann mit statistischen Verfahren und jenen aus der Informationstheorie ermittelt werden. Die *„Aktualität"* sollte die Zeitstempelqualität abbilden. Die Neudefinition lautet daher:

„Daten sind aktuell, wenn die Zeitspanne zwischen der Datenerzeugung und Datenaufzeichnung oder –speicherung minimal ist."

Die Zeitstempelqualität ist für Prognoseverfahren oder die Ursache-Wirkungssuche ausschlaggebend. Die Vollständigkeit muss nicht neu definiert werden, auch wenn der Bezug auf den festgelegten Zeitpunkt in gewisser Weise durch die neue Definition der Aktualität überbestimmt ist. Der „*angemessene Umfang*" bedarf ebenfalls keiner Neudefinition, wenngleich festzuhalten ist, dass ein Zuviel an Daten im Big Data Umfeld nicht mit einer schlechten Datenqualität gleichgesetzt wird. Die Definition der Wertschöpfung bleibt unverändert. Sie wird im Kontext dieser Arbeit jedoch nicht weiter betrachtet.

Die inhärenten Datenqualitätsdimensionen werden von vier auf zwei reduziert. Die „*Fehlerfreiheit*" bleibt in dieser Datenqualitätskategorie unverändert. Die „*Objektivität*" wird aus der betrachteten Gesamtheit der Datenqualitätsdimensionen entfernt, da sie für industrielle Analyseanwendungen nicht relevant ist. Die „*Glaubwürdigkeit*" und das „*hohe Ansehen*" der Daten wird in folgender Definition der „*Glaubwürdigkeit*" zusammengeführt:

„Daten sind glaubwürdig, wenn deren Gewinnung und Verbreitung mit hohem Aufwand betrieben werden und die dafür eingesetzten Systeme hohes Ansehen haben."

Die automatische, vom Menschen unabhängige Datenerfassung hat das höchste Ansehen. Abbildung 39 und deren Beschreibung gehen näher darauf ein.

Von den ursprünglichen 15 Datenqualitätsdimensionen sind für den industriellen datengestützten Analyseprozess zehn relevant. Wo nötig wurde die Definition angepasst, bzw. eine neue Definition vorgenommen. Diese zehn Datenqualitätsdimensionen werden im Reifegradmodell ihre Anwendung finden, um die Datenqualität erstmals im Umfeld der Big Data Analytik zu messen.

Das Kapitel ist von Relevanz, da es die Wichtigkeit des Daten- und Informationsmanagements zeigt. Erkenntnisse daraus werden in das Reifegradmodell übernommen. So ist für die Speicherung und Verteilung von Daten ein Data-Warehouse ein System von hoher Reife und als solches zu bevorzugen.

Das Datenqualitätsmanagement soll im übertragenen Sinne seine Anwendung finden, indem das Reifegradmodell, mit der Empfehlung zur Verbesserung in den Reifegradkategorien, zur Verbesserung und Stabilisierung des Datenqualitätsniveaus beitragen soll.

Folgend dem *fit for use* Ansatz, soll das Reifegradmodell die Reife der Daten und des Systems für einen Anwendungsfall messen. Dieser Anwendungsfall ist der datenanalytische Prozess. Im folgenden Kapitel werden die Grundlagen der Datenanalyse genauer beschrieben.

4 Datenanalytische Grundlagen

Der Begriff Data-Mining[261] wurde in den 1980er Jahren zum ersten Mal von LOVELL verwendet, um damit unterschiedliche Verfahren zur Auswertung von Daten, der automatischen oder mathematisch unterstützten Auswahl von Daten und der Modellspezifikation durch Daten zusammenzufassen.[262] Eine allgemein gültige Definition wurde hierzu nicht gegeben.

FAYYAD ET AL definieren Data Mining in ihrem KDD[263]-Prozess folgend:

„Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data."[264] Data-Mining wird als die ausschließliche Anwendung von Algorithmen und statistischer Verfahren auf die Daten angesehen.

WIRTH UND HIPP definieren Data-Mining im Rahmen des CRISP-DM[265] als „[…] a creative process which requires a number of different skills and knowledge."[266] Diese umfassendere Sichtweise hat sich in den letzten Jahren durchgesetzt.

WITTEN UND FRANK definieren Data Mining als „[…] techniques for finding and describing structural patterns in data as a tool for helping to explain that data and make predictions from it."[267]

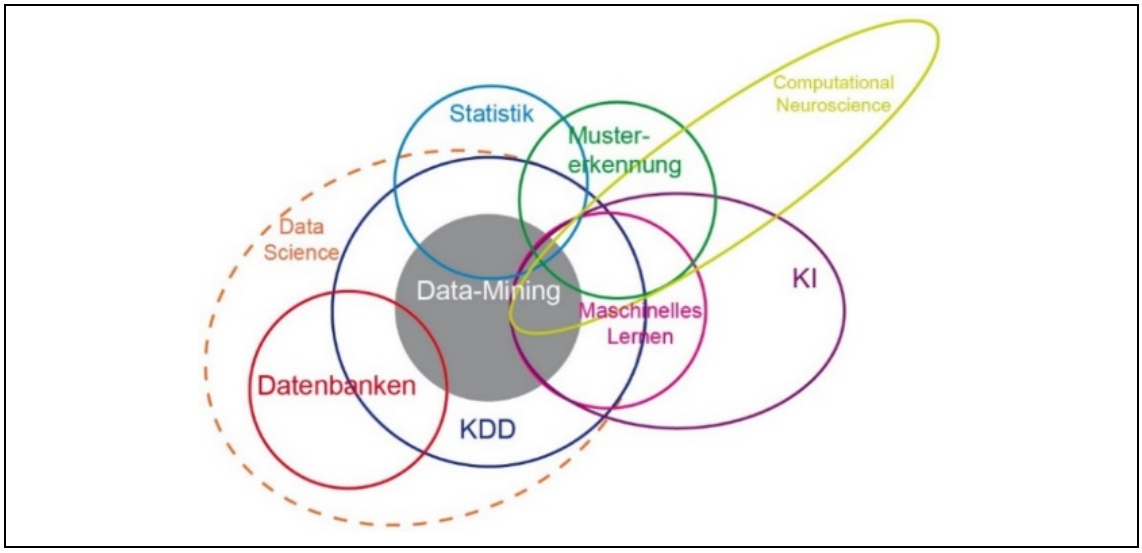

Abbildung 16: Multidisziplinäres Feld des Data-Mining[268]

[261] Data Mining im Original
[262] Vgl. Lovell, M. C. (1983), S. 1.
[263] KDD: Knowledge Discovery in Databases
[264] Vgl. Fayyad, U. et al. (1996), S. 40.
[265] CRISP-DM: Cross-Industry Standard Process for Data Mining
[266] Wirth, R.; Hipp, J. (2000), S. 29.
[267] Witten, I. H.; Frank, E. (2005), S. 9.
[268] Quelle: Hall, P. et al. (2014), S. 3. (aus dem Englischen übersetzt)

RUNKLER definiert Data-Mining als einen Prozess zur Generierung oder Extrahierung von Wissen aus Daten. Das Wissen manifestiert sich als interessante Muster, welche allgemein gültig, nicht trivial, neu, nützlich und verständlich sind.[269]

Diese generische Definition von Data-Mining wird in dieser Arbeit verwendet. Abbildung 16 zeigt diese Sicht auf Data-Mining, indem Data-Mining im Zentrum der Vielfalt der Datenanalyse steht und sich dieser bedient, anstatt ein isolierter Schritt in einem Vorgehen zu sein. So ist Wissen über Datenbanken mit den ETL- und OLAP-Funktionalitäten von Data-Warehouses ein wichtiger Vorverarbeitungsschritt. Maschinelles Lernen umfasst Methoden und Algorithmen, die über die reine Statistik hinausgehen. In Kombination können die beiden Felder zur Mustererkennung eingesetzt werden, welche wiederum eine wesentliche Aufgabe von Data-Mining ist.

4.1 Generische Modelle zur Datenanalyse

Die methodische Analyse von Daten ist ein Prozess der mehr als die Anwendung von Algorithmen umfasst. Er erstreckt sich von der Datenaufnahme und -auswahl bis zur Evaluierung der Ergebnisse und der Implementierung in den Geschäftsprozess.

Zur Standardisierung des Vorgehens, wurden ab den 1990er Jahren unterschiedliche Referenzmodelle entwickelt. Folgend wird ein Auszug daraus kurz vorgestellt, wobei auf das CRISP-DM Modell im Detail eingegangen wird.

4.1.1 CRISP-DM Modell

Einer der am weitesten verbreiteten Prozesse ist dabei der CRISP-DM[270] welcher von einem Konsortium aus NCR[271], SPSS[272], OHRA[273] und DaimlerChrysler entwickelt wurde. Abbildung 17 zeigt die sechs Phasen des Prozesses, wobei Data-Mining kein eigener Schritt ist, sondern der gesamte Prozess als Data-Mining gesehen wird.

Ziel von CRISP-DM war es ein Vorgehen zu schaffen, welches unabhängig von der eingesetzten Industrie, der verwendeten Werkzeuge und Methoden und der Zielstellung ist. Der entwickelte Standard sollte die Bearbeitung von datenanalytischen Projekten in einem vereinheitlichten Rahmen ermöglichen und die Erfolgsgarantie erhöhen.[274]

CRISP-DM besteht aus einem Referenz- und einem Prozessmodell. Das Referenzmodell definiert die Phasen eines Data-Mining Projektes. Jede Phase hat generische Handlungsfelder, die in spezifische Handlungen und in weiterer Folge Prozessschritte unterteilt werden, abhängig vom Einsatzgebiet des Data-Mining Projektes. Das Vorgehen zur Auswahl der spezifischen Handlungsfelder und der

[269] Vgl. Runkler, T. A. (2010), S. 2.
[270] CRISP-DM: Cross Industry Standard Process for Data Mining
[271] NCR Systems Engineering Copenhagen (USA und Dänemark)
[272] SPSS Inc. (Softwarehaus aus den USA, später gekauft von IBM)
[273] OHRA Verzekeringen en Bank Groep B.V. (Niederlande)
[274] Vgl. Chapman, P. et al. (2000), S. 1.

jeweiligen Prozessschritte, sowie deren genaue Ausgestaltung sind im Prozessmodell von CRISP-DM beschrieben.[275]

Im Vergleich zum früher veröffentlichten KDD-Prozess, der direkt mit der Datensammlung beginnt, startet der CRISP-DM mit dem Verständnis des Projektziels und der Daten an sich.[276]

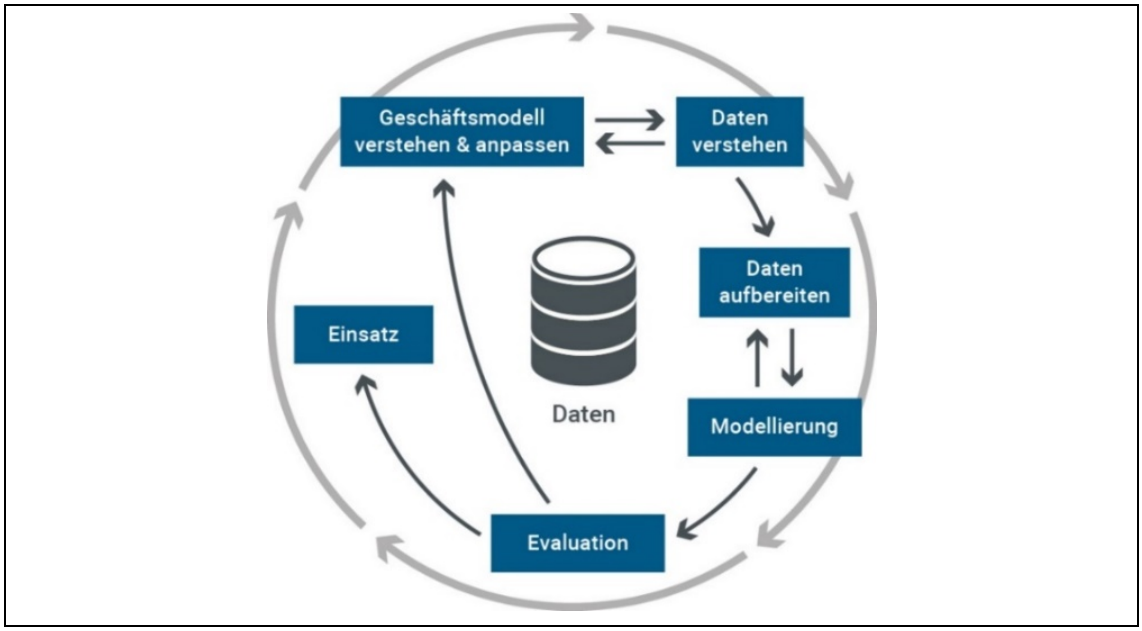

Abbildung 17: CRISP-DM Referenzmodell[277]

Die sechs Phasen des Referenzmodells müssen nicht zwingend in der gegebenen Reihenfolge durchlaufen werden. Die Pfeile zwischen den Phasen stellen die üblichste Interaktion dar. Iterationsschleifen zwischen jeder Phase sind jedoch üblich. Der graue Pfeilkreis symbolisiert, dass man aus jedem Projekt im Data-Mining lernt und die gewonnenen Erfahrungen neue Fragen aufwerfen und Projekte starten können.[278]

Geschäftsmodell verstehen (Business Understanding):

Hier gilt es die Ziele des Projektes zu verstehen. Der Fokus liegt auf der Sicht des Auftraggebers und der generellen Problemstellung. In weiterer Folge wird diese Problemstellung in ein Data-Mining Problem übersetzt und ein Projektplan erstellt.[279]

Unternehmens- und Geschäftsziele festlegen (determine business objectives)

Der Datenanalyst muss sich mit den Zielen des Vorhabens vertraut machen, um zu verstehen was erreicht werden will. Der Geschäftshintergrund (*background*) fasst die Situation des Unternehmens am Beginn des Projektes zusammen. In den Geschäftszielen (*business objectives*) werden die Haupt- und Nebenziele des Vorhabens aus Sicht des Unternehmens und des Auftraggebers definiert. Die

[275] Vgl. Chapman, P. et al. (2000), S. 6 f.
[276] Vgl. Wirth, R.; Hipp, J. (2000), S. 30.
[277] Quelle: in Anlehnung an Wirth, R.; Hipp, J. (2000), S. 33.
[278] Vgl. Wirth, R.; Hipp, J. (2000), S. 32.
[279] Vgl. Chapman, P. et al. (2000), S. 10.

Erfolgsfaktoren (*business success criteria*) sind jene Kriterien, die erfüllt werden müssen, damit das Projekt als Erfolg betrachtet wird.[280]

Umfeld bewerten (assess situation)

Es wird erhoben, welche Ressourcen, Hindernisse und Voraussetzungen bei der späteren Definition des Data-Mining Ziels betrachtet werden müssen. Die Analyse ist umfangreich und umfasst vier Outputs. Die Inventarliste über die Ressourcen (*inventory of resources*) listet alle Ressourcen auf, die dem Projekt zur Verfügung stehen. Dazu zählen Personal, Daten-, Hardware- und Softwareressourcen. In der Aufzählung zu den Anforderungen, Voraussetzungen und Hemmnissen (*requirements, assumptions, and constraints*) wird erfasst was für die Projektdurchführung bereitgestellt werden muss, wie z. B. Zugangsdaten. Des Weiteren sind Voraussetzungen an die Daten zu formulieren, um das Data-Mining durchführen zu können. Die Hemmnisse beschreiben organisatorische die sich auf die Ressourcen auswirken sowie technische, die sich auf den Datenumfang auswirken können. Die Liste der Risiken und Handlungsalternativen (*risks and contingencies*) beschreibt jene Risiken, die das Projekt verzögern oder scheitern lassen können und zeigt Handlungsalternativen auf, um den Risiken zu begegnen. Die speziellen Terminologien (*terminology*) werden in ein Glossar zusammengefasst. Das letzte Ergebnis der Umfeldbewertung umfasst eine Kosten-Nutzen-Analyse (*costs and benefits*).[281]

Data-Mining Ziele festlegen (determine data mining goals)

Nachdem die Haupt- und Nebenziele des Vorhabens festgelegt wurden, werden diese vom Data Scientist in Data-Mining Ziele (*data mining goals*) übersetzt. Die Erfolgsfaktoren für das Data-Mining (*data mining success criteria*) sind technische Größen wie die Prognosegenauigkeit oder der Lift bei Assoziationsanalysen.[282]

Projektplan erstellen (produce project plan)

Es wird bestimmt wie die formulierten Data-Mining Ziele erreicht werden sollten. Der Projektplan (*project plan*), beinhaltet alle Schritte des Projektes, mit den zeitlichen Angaben und den benötigten Ressourcen, Eingangsgrößen und Ergebnissen sowie den gegenseitigen Abhängigkeiten. Der Projektplan ist dynamisch und sollte nach jeder Phase auf Abweichungen hin evaluiert und wenn nötig angepasst werden. Das zweite Ergebnis ist eine erste Bewertung der Werkzeuge, Methoden und Techniken (*initial assessment of tools and techniques*).[283]

Daten verstehen (Data Unterstanding):

Zur ausreichenden Beschreibung des Data-Mining Problems und zur Erstellung eines Projektplans ist ein Verständnis der zur Verfügung stehenden Daten notwendig.[284] In dieser Phase werden Daten gesammelt und Methoden eingesetzt, um ein besseres

[280] Vgl. Chapman, P. et al. (2000), S. 14.
[281] Vgl. Chapman, P. et al. (2000), S. 14 f.
[282] Vgl. Chapman, P. et al. (2000), S. 16.
[283] Vgl. Chapman, P. et al. (2000), S. 16 f.
[284] Vgl. Wirth, R.; Hipp, J. (2000), S. 34.

Verständnis für die Daten zu erlangen. Es werden ggf. erste Probleme mit der Datenqualität erkannt und beschrieben und wenn möglich erste Hypothesen basierend auf interessanten Subsets von Daten geformt.[285]

Ausgangsdaten sammeln (collect initial data)

Jene Daten, welche in der ersten CRISP-DM Phase beschrieben wurden, werden zugänglich gemacht. Der Bericht über die Ausgangsdaten (*initial data collection report*) ist der Output des Schrittes. Er umfasst den Speicherort, die Zugangsmethode und etwaige Probleme beim Zugriff auf die Daten.[286] Sollte die Ausgangsmenge aus einer Zusammenführung mehrere Datenquellen entstehen, ist darauf zu achten, dass dadurch Datenqualitätsprobleme entstehen können. Gründe hierführt können Inkonsistenzen durch z. B. Zeitstempelprobleme sein. Dieser Schritt umfasst in einem kleinen Umfang die Datensammlung für das gesamte Projekt.[287]

Daten beschreiben (describe data)

Es werden die wichtigsten Eigenschaften über die Daten in einem Bericht (*data description report*) zusammengefasst. Er gibt Auskunft über das Datenformat, den Umfang der Daten, das Skalenniveau und die Bedeutung der Spalten. Der Bericht beurteilt, ob die Daten die Anforderungen für das Projektziel erfüllen.[288]

Daten untersuchen (explore data)

Die Daten werden visualisiert oder mit statistischen Verfahren untersucht, um erste Muster und Zusammenhänge untereinander und zum Data-Mining Ziel zu finden. Der resultierende Bericht (*data exploration report*) enthält mögliche Funde und daraus folgende Hypothesen und deren Bedeutung für das Data-Mining Projekt.[289]

Datenqualität prüfen (verify data quality)

Die Datenqualität wird bezogen auf die Zielerreichung beurteilt[290]. Es müssen alle Fälle der Fragestellung des Projektes in den Daten abgebildet sowie Fehlstellen und Fehler dokumentiert werden. Der Datenqualitätsbericht (*data quality report*), als Prozessschrittergebnis, enthält mögliche Probleme und Lösungsvorschläge.[291]

Daten aufbereiten (Data Preparation)

In dieser Phase wird der finale Datensatz (*dataset*) erzeugt, der in der Modellierungsphase verwendet wird. Die Schritte umfassen die Auswahl von Datentabellen und Attributen, die Beseitigung von Fehlstellen und Ausreißern, die Erzeugung neuer Attribute und die Transformation von Skalenniveaus passend für die

[285] Vgl. Chapman, P. et al. (2000), S. 10.
[286] Vgl. Chapman, P. et al. (2000), S. 18.
[287] Vgl. Chapman, P. et al. (2000), S. 38 f.
[288] Vgl. Chapman, P. et al. (2000), S. 18.
[289] Vgl. Chapman, P. et al. (2000), S. 18 f.
[290] Man spricht in diesem Fall von „fit for use“
[291] Vgl. Chapman, P. et al. (2000), S. 19.

Modellierungsmethoden.[292] Die Beschreibung des Datensatzes (*dataset description*) ist neben dem Datensatz der zweite globale Output der Phase. Die Verteilung, das Skalenniveau und andere Besonderheiten werden dokumentiert.[293]

Daten auswählen (select data)

Die Entscheidung welche Daten ausgewählt werden hängt von Faktoren, wie der Relevanz für das Data-Mining Ziel, der Qualität der Daten, der Datentypen und -formate und des Umfangs der Daten ab. Es werden die Attribute und Datensätze gewählt. Das Ergebnis dieses Schrittes umfasst eine Beurteilung über die Aufnahme oder den Ausschluss der Daten für die Auswahl (*rationale for inclusion/exclusion*) mit einer Begründung für die Entscheidung.[294]

Daten bereinigen (clean data)

Hier werden jene Aktionen gesetzt, die nötig sind, um die Datenqualität auf ein Niveau zu heben, dass für den Analysealgorithmus verarbeitbar ist. Hauptaufgabe ist die Entfernung von Ausreißern und die Ersetzung von Fehlstellen. Ein umfassender Bericht zu den Maßnahmen der Datenbereinigung (*data cleaning report*) ist das Ergebnis des Schrittes.[295]

Daten erstellen (construct data)

In diesem Schritt werden neue Attribute oder Datensätze erzeugt. Diese abgeleiteten Attribute (*derived attributes*) sind ein wesentliches Ergebnis des Schrittes. Das zweite Ergebnis sind die neuen Datensätze. Deren Generierung (Vorgehensweise und eingesetzte Methoden) muss detailliert beschrieben werden.[296]

Datenintegration (integrate data)

Daten aus unterschiedlichen Quellen und Tabellen werden zu einer neuen umfassenden Quelle zusammengeführt. Dadurch erhöht sich der Informationsgehalt für einen Datensatz, da die Quellen neue Daten über den Datensatz enthalten. Das Ergebnis des Schrittes sind die zusammengeführten Daten (*merged data*).[297]

Daten formatieren (format data)

Die Maßnahmen sind vordergründlich syntaktischer Natur. Modellierungsverfahren benötigen gewisse Skalenniveaus und die Transformation in diese erfolgt in diesem Schritt. Die neu formatierten Daten (*reformatted data*) sind das Ergebnis dieses Schrittes.[298]

[292] Vgl. Chapman, P. et al. (2000), S. 11.
[293] Vgl. Chapman, P. et al. (2000), S. 21.
[294] Vgl. Chapman, P. et al. (2000), S. 21.
[295] Vgl. Chapman, P. et al. (2000), S. 21.
[296] Vgl. Chapman, P. et al. (2000), S. 21 f.
[297] Vgl. Chapman, P. et al. (2000), S. 22.
[298] Vgl. Chapman, P. et al. (2000), S. 22 f.

Modellierung (Modeling)

In dieser Phase werden passende Modellierungsverfahren auf den finalen Datensatz angewandt. Für einen Datensatz und ein Problem kommen in vielen Fällen mehrere Modellierungsverfahren in Frage. Ziel der Modellierungsphase ist es die Parameter jedes Verfahrens optimal einzustellen.[299]

Da die Verfahren unterschiedliche Anforderungen an die Formate der Daten haben, kann es vermehrt zu Iterationsschleifen zwischen den Phasen *Daten aufbereiten* und *Modellierung* kommen. Des Weiteren ist es möglich, dass Probleme in den Daten erst in der Modellierungsphase erkannt werden.[300]

Auswahl des Modellierungsverfahrens (select modeling technique)

Die Erstauswahl passiert bereits in der ersten Phase des CRISP-DM. In der Modellierungsphase fließen jedoch die zusätzlichen Informationen der weiteren Phasen ein. Die Ergebnisse dieses Teilschrittes sind eine genaue Beschreibung des Verfahrens *(modeling technique*) und die Dokumentation der Voraussetzungen, die das Verfahren an die Daten stellt. (*modeling assumptions*). Solche Voraussetzungen können die nötige Verteilungsform der Daten sein, das Skalenniveau oder dass es keine fehlenden Werte geben darf.[301]

Erzeuge ein Testdesign (generate test design)

Das Testdesign dient dazu, das Modell zu validieren. Abhängig von der datenanalytischen Methode muss das Testdesign spezifische Anforderungen erfüllen. Als Ergebnis dieses Schrittes wird das Testdesign (*test design*) erzeugt und beschrieben, wie der Datensatz für Training, Test und Validierung geteilt wird.[302]

Modell erstellen (build model)

Das Modell oder die Modelle werden mit der Software erstellt. Modelle und die Softwareoberflächen bieten die Möglichkeit mittels Parameter Einstellungen vorzunehmen. Diese Parameterwerte (*parameter settings*) müssen dokumentiert werden. Die Modelle selbst (*Models*) sind das zweite Ergebnis dieses Schrittes. Hierbei handelt es sich nur um das Modell. Der spezifische Bericht ist das dritte Ergebnis (*model description*).[303]

Modell bewerten (assess model)

Der Data Scientist interpretiert das Modell entsprechend seines Domänenwissens, der Erfolgsfaktoren und des Testdesigns. Seine Einschätzung ist technischer Natur. Domänenexperten geben Input zu den Ergebnissen in einen Geschäftskontext. Basierend auf diesen Rückmeldungen werden die Modelle vom Data Scientist gereiht.

[299] Vgl. Chapman, P. et al. (2000), S. 11.
[300] Vgl. Wirth, R.; Hipp, J. (2000), S. 34.
[301] Vgl. Chapman, P. et al. (2000), S. 24.
[302] Vgl. Chapman, P. et al. (2000), S. 24.
[303] Vgl. Chapman, P. et al. (2000), S. 24 f.

Die Ergebnisse umfassen eine Modellbewertung (*model assessment*) und eine Überarbeitung der Parametereinstellungen (*revised parameter settings*).[304]

Evaluation (Evaluation)

Am Ende der Modellierungsphase liegen meist mehrere Modelle vor. In der Evaluationsphase wird jenes Modell gewählt, welches das beste Ergebnis liefert. Dabei ist die Qualität des Ergebnisses aus analytischer Sicht allein nicht ausschlaggebend. Es muss evaluiert werden, ob alle Ziele aus der ersten Phase erreicht wurden. Am Ende dieser Phase wird darüber entschieden, ob die Ergebnisse verwendet werden können.[305]

Evaluiere Ergebnisse (evaluate results)

Es wird festgestellt wie die Zielerreichung bezogen auf die Unternehmensziele des Data-Mining Projektes ist. Das finale Assessment der Ergebnisse (*assessment of data mining results with respect to business success criteria*) mit einer Einschätzung, ob die Unternehmensziele erreicht wurden und ein Einsatz möglich ist, ist das erste Ergebnis des Prozessschrittes. Danach werden jene Modelle freigegeben (*approved models*), welche die Erfolgsfaktoren erfüllen. Sie stellen das zweite Ergebnis des Prozessschrittes dar.[306]

Prozess bewerten (review process)

Falls die Ergebnisse zufriedenstellend sind, wird der Prozess des Data-Minings genauer überprüft. Der Fokus liegt auf Punkten, die unter Umständen übersehen wurden und auf der Qualitätssicherung, die sicherstellen soll, dass das Modell korrekt erstellt wurde. Die Erkenntnisse werden in einem Bericht zusammengefasst (*review of process*).[307]

Nächste Schritte festlegen (determine next steps)

Basierend auf den Ergebnissen der vorhergehenden Schritte, wird festgelegt, ob das Verfahren implementiert, der Prozess an einer früheren Phase neu gestartet wird oder ob das Projekt abgebrochen werden muss. Die Ergebnisse sind eine Liste der möglichen Handlungen (*list of possible actions*) und eine Beschreibung der Entscheidung (*decision*).[308]

Einsatz (Deployment)

Das erzeugte Wissen aus den Daten im Zuge des Data-Mining Projektes muss in einer für den Auftraggeber geeigneten Form dargestellt werden. Abhängig von den Anforderungen ist die Phase schnell umgesetzt, wenn es um eine einfache Implementierung von Entscheidungsunterstützungen geht oder aufwendiger, wenn ein laufender Data-Mining Prozess unternehmensübergreifend implementiert werden soll.[309] Der Aufwand der Implementierung und die generelle Möglichkeit selbiger, bei z. B. einer

[304] Vgl. Chapman, P. et al. (2000), S. 25.
[305] Vgl. Chapman, P. et al. (2000), S. 11.
[306] Vgl. Chapman, P. et al. (2000), S. 26 f.
[307] Vgl. Chapman, P. et al. (2000), S. 27.
[308] Vgl. Chapman, P. et al. (2000), S. 27.
[309] Vgl. Chapman, P. et al. (2000), S. 11.; Wirth, R.; Hipp, J. (2000), S. 34 f.

automatischen Sammlung und Auswertung der Daten, sollte bereits in der ersten Phase des CRISP-DM beachtet werden.

Einsatz planen (plan deployment)

Um den Einsatz des Ergebnisses im Rahmen des operativen Geschäfts zu planen wird ein knapper Projektplan (*deployment plan*) erstellt. Er enthält die nötigen Schritte, um das finale Modell und die Subergebnisse in den regulären Betrieb zu überführen.[310] Ein wesentlicher Punkt ist eine Vorgehensweise die Informationen an die Endbenutzer weiterzugeben.[311]

Plane Überwachung und Wartung (plan monitoring and maintenance)

Um zu vermeiden, dass im täglichen Einsatz über die Zeit falsche Ergebnisse erzeugt werden, muss eine Strategie zu Überwachung der korrekten Anwendung des Modells implementiert werden. Sie beinhaltet Schritte zur Wartung und Anpassung bei Abweichungen. Die nötigen Schritte sind in einem Vorgehensplan (*monitoring and maintenance plan*) festzuhalten.[312]

Endbericht erstellen (produce final report)

Am Ende des Projektes werden alle beschriebenen Ergebnisse der vorherigen Phasen und deren Prozessschritte in einem Endbericht (*final report*) zusammengeschrieben. Die wichtigsten Ergebnisse und Rückschlüsse für neue Projekte können in einer Endpräsentation (*final presentation*) für den Kunden kurz zusammengefasst werden.[313]

Projekt prüfen und bewerten (review project)

Im letzten Prozessschritt des gesamten CRISP-DM werden die Erfahrungen zusammengeschrieben (*experience documentation*), die mit dem Ablauf des Projekts zu tun haben. Dazu zählen Hindernisse, falsche Herangehensweisen und Hinweise, um bei ähnlichen Problemen schnell die richtigen Daten und Modellierungsmethoden zu wählen.[314] Ein Teil der Bewertung ist die Kosten-Nutzen Analyse basierend auf den Inputs der ersten CRISP-DM Phase.[315]

4.1.2 Weitere Prozessmodelle

Neben dem CRISP-DM gibt es noch weitere Prozessmodelle für die Datenanalyse. Sie wurden von facheinschlägigen Softwareanbietern und Wissenschaftlern entwickeln, um bestimmte Spezialaspekte abzubilden oder den Prozess neuen Entwicklungen anzupassen. MARISCAL ET AL. geben einen Überblick über diese Modelle und zeigen wie sie zueinanderstehen.[316]

[310] Vgl. Chapman, P. et al. (2000), S. 28.
[311] Vgl. Chapman, P. et al. (2000), S. 54.
[312] Vgl. Chapman, P. et al. (2000), S. 29.
[313] Vgl. Chapman, P. et al. (2000), S. 29.
[314] Vgl. Chapman, P. et al. (2000), S. 29.
[315] Vgl. Chapman, P. et al. (2000), S. 63
[316] Vgl. Mariscal, G. et al. (2010), S. 142.

Knowledge Discovery in Databases (KDD)

FAYYAD veröffentlichte 1996 mit dem KDD-Prozess das erste Modell zur standardisierten Datenanalyse. Es besteht aus fünf Schritten die den Analyseprozess von der Datenauswahl, über das Data-Mining bis zur Ergebnisinterpretation beschreiben. Abbildung 18 zeigt die Prozessschritte und die Ergebnisse der Prozesse.[317]

Sollte am Ende des Evaluationsschrittes das Ergebnis nicht zufriedenstellend sein, muss der Prozess an einer passenden Stelle, mit allen folgenden Prozessschritten wiederholt werden.[318] Die Prozessschritte, mit Ausnahme von Data-Mining, und deren Ergebnisse werden nicht wie beim CRISP-DM genauer beschrieben. Beim Prozessschritt Data-Mining werden die unterschiedlichen Methoden, anhand eines Beispiels erörtert.[319]

Der KDD-Prozess behandelt sehr generisch die Wahl, Vorverarbeitung, Transformation, Analyse und Interpretation von Daten. Eine genauere Definition von Subprozessschritten wird nicht gegeben. Des Weiteren werden wesentliche Schritte zur Projektplanung und des generellen Verständnisses der Daten und der folgenden Implementierung des Ergebnisses nicht behandelt.

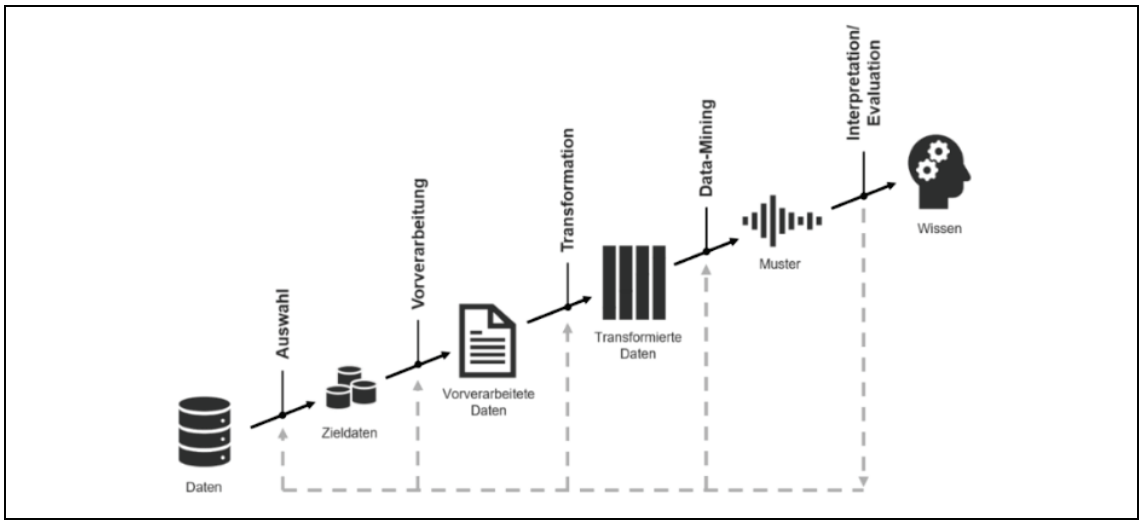

Abbildung 18: KDD-Prozess[320]

Knowledge Discovery in Industrial Databases (KDID)

LIEBER ET AL. entwickelten das KDID-Modell, um Kritikpunkten an den klassischen Prozessmodellen zu begegnen. Diese gehen davon aus, dass große vereinte Datenbestände bereits vorliegen. Im industriellen Umfeld ist dieser Umstand nicht immer gegeben. Die Daten sind auf unterschiedliche Systeme verteilt und werden in unterschiedlichen Zeithorizonten erfasst, von einer Abtastung im Millisekundenbereich bei Sensoren bis hin zu tagesgenauen Daten beim Planungsprozess.[321]

Der KDID versucht die Basis des KDD-Prozesses mit den Stärken des CRISP-DM Prozesses zu verbinden. Abbildung 19 zeigt den Ablauf der Wissensentdeckung mit dem

[317] Vgl. Fayyad, U. et al. (1996), S. 40 f.
[318] Vgl. Fayyad, U. et al. (1996), S. 42.
[319] Vgl. Fayyad, U. et al. (1996), S. 43 ff.
[320] Quelle: Fayyad, U. et al. (1996), S. 41. (aus dem Englischen übersetzt)
[321] Vgl. Lieber, D. et al. (2013), S. 389 f.

KDID-Modell. Der spezielle Fokus des Modells liegt auf der genauen Zieldefinition mit dem Kunden und dessen Experten auf dem Einsatzgebiet. Weitere Schnittpunkte mit dem CRISP-DM sind die Abbildung des IST-Zustandes der IT-Struktur und des Prozesses, der Herausforderungen der Datenbereitstellung im industriellen Umfeld und dem Einsatz der Ergebnisse in operativen Betrieb.

Den Herausforderungen in der Datenbereitstellung wird mit einem zweistufigen Prozessschritt zur Datensammlung begegnet. Die Datensammlung teilt sich in die Datenintegration und Datenerfassung und –speicherung. Das Ziel ist die Bedeutung einer durchgehenden Datenerfassung und strukturierten Datenhaltung hervorzustreichen.[322]

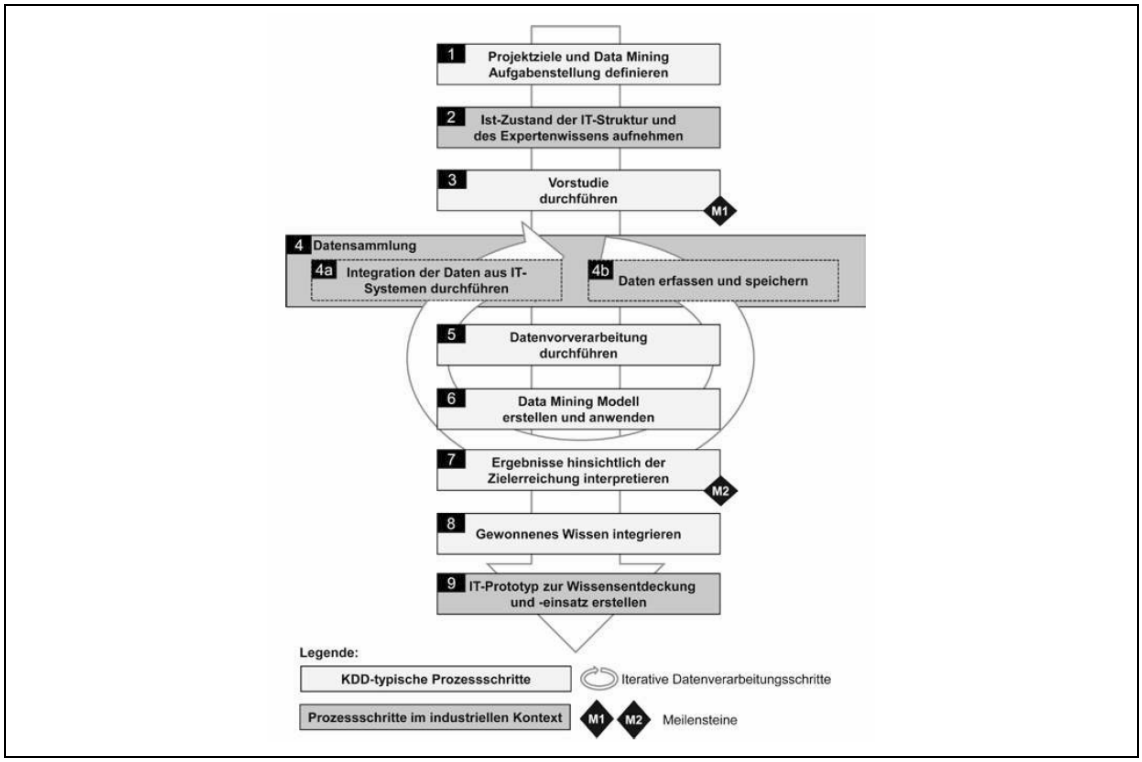

Abbildung 19: KDID-Prozess[323]

Sample, Explore, Modify, Model, Assess (SEMMA)

SEMMA wurde vom SAS Institute[324] veröffentlicht, um die Analyseschritte des eigenen Softwaretools, dem SAS Enterprise Miner, als Prozessmodell abzubilden. SEMMA steigt ohne den expliziten Schritt des Verständnisses des Geschäftsumfeldes und der Daten in die Analyse ein. Es existiert des Weiteren kein Schritt für die Integration des entdeckten Wissens in die Geschäftstätigkeit des Unternehmens und es sind keine Iterationen vorgesehen, um Prozessschritte zu wiederholen.[325]

[322] Vgl. Lieber, D. et al. (2013), S. 390 ff.
[323] Quelle: Lieber, D. et al. (2013), S. 390.
[324] SAS Institute: Softwarehaus und Anbieter von Business Analytics Tools. SAS steht für Statistical Analysis System.
[325] Vgl. Mariscal, G. et al. (2010), S. 144.

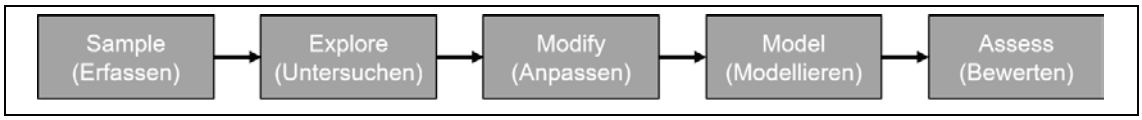

Abbildung 20: SEMMA Ablauf[326]

Data Mining and Knowledge Discovery (DMKD) nach Cios und Kurgan

CIOS ET AL. veröffentlichten ihr Modell mit einem verstärkten Forschungsfokus mit dem finalen Schritt, das erzeugte Wissen in eine Vielzahl von Domänen zu transferieren und nicht alleine in jene des Einsatzbeispiels. Es basiert auf dem CRISP-DM Modell, detailliert jedoch die Iterationsschleifen genauer.[327]

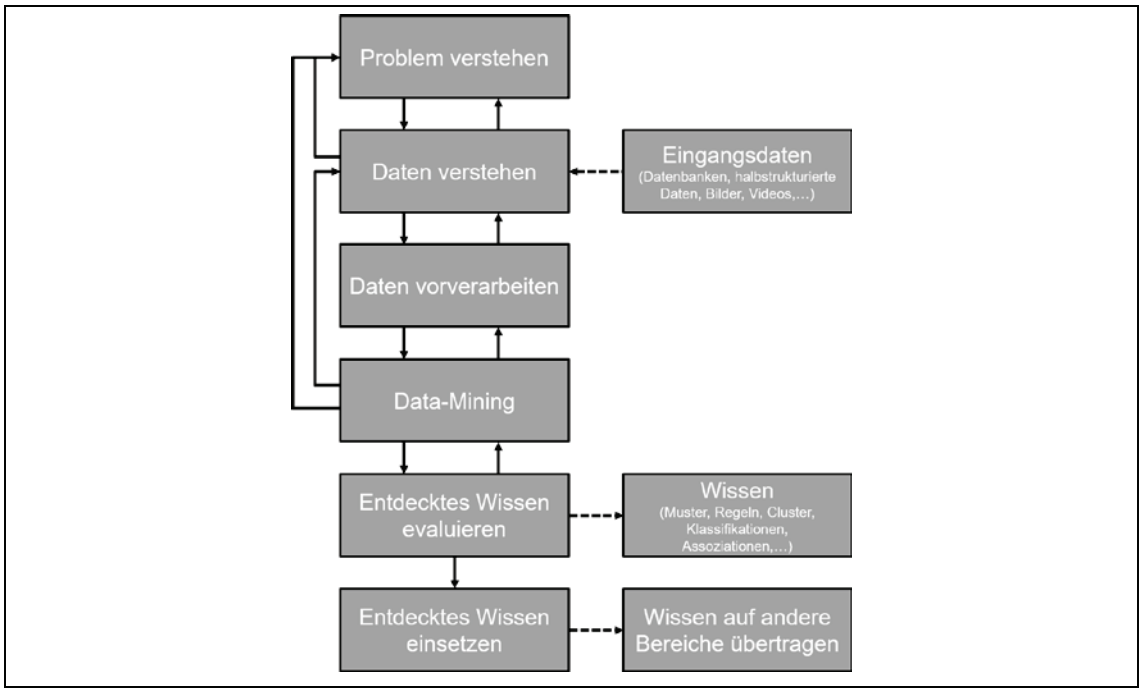

Abbildung 21: DMKD nach CIOS UND KURGAN[328]

Die beiden letzten Schritte sind markante Unterschiede zu anderen Prozessmodellen für die Datenanalyse. Das entdeckte Wissen wird mit Domänenexperten evaluiert. Sie helfen die Ergebnisse in Kontext zu setzen und den Einfluss im Einsatzgebiet abzuschätzen. Dabei werden nur in diesem Schritt freigegebene Modelle eingesetzt und der Prozess der Generierung wird für zukünftige Projekte betrachtet und gewonnene Erkenntnisse werden dokumentiert. Der letzte Schritt legt fest, wie das neue Wissen implementiert werden soll und auf andere Bereiche übertragen werden kann.[329]

4.2 Analysealgorithmen

Data-Mining umfasst ein breites Spektrum an Methoden zur Analyse von großen Datenmengen. Sie werden aus einigen Feldern von Abbildung 16, wie der Statistik, des

[326] Quelle: in Anlehnung an Mariscal, G. et al. (2010), S. 144.
[327] Vgl. Cios, K. J.; Kurgan, Lukasz A. (2004), S. 6.
[328] Quelle: in Anlehnung an Cios, K. J. et al. (2007), S. 15.
[329] Vgl. Cios, K. J.; Kurgan, Lukasz A. (2004), S. 8.

maschinellen Lernens und der Mustererkennung entnommen. Die Methoden können abhängig von deren Zielsetzung oder des grundlegenden Analyseansatzes gegliedert werden.

4.2.1 Einteilung der Analysemethoden

Bei der Analyse großer Datenmengen kommt die multivariate Statistik zur Anwendung. Multivariate Verfahren untersuchen Datensätze mit mehr als einer Variablen auf Zusammenhänge. Würde nur eine Variable – auch Attribut genannt – existieren, handelt es sich um ein univariates Beispiel.[330]

Die Statistik unterscheidet die beiden relevanten Felder der strukturprüfenden und strukturentdeckenden Methoden. Die deskriptiven Methoden seinen der Vollständigkeit halber erwähnt, obgleich sie bei der Analyse großer Datenmengen eine untergeordnete Rolle spielen. Strukturprüfende Verfahren prüfen Zusammenhänge in Datensätzen. Dabei existiert a priori bereits eine Hypothese über die kausalen Zusammenhänge. Die abhängige Variable und die unabhängigen Variablen in der Problemstellung sind bekannt. Die Verfahren der Strukturprüfung sind u. a. die Regressionsanalyse, Zeitreihenanalyse, Varianzanalyse oder Diskriminanzanalyse. Strukturentdeckende Verfahren setzen kein Wissen über die Abhängigkeit der Variablen voraus. Deren primäre Aufgabe ist es Zusammenhänge in den Datensätzen, ohne vorher existierende Hypothesen zu finden. Die Clusteranalyse oder die Faktorenanalyse sind zwei der Verfahren, die bei der Strukturentdeckung zu Anwendung kommen.[331]

Das maschinelle Lernen nimmt eine Gliederung in überwachte und unüberwachte Verfahren vor. Teilüberwachte Verfahren[332] stellen eine Kombination der beiden dar. Bestärkende Lernverfahren[333] sind ein Sonderfall welche verstärkt in der Robotik eingesetzt werden.

Überwachte Lernverfahren funktionieren unter der Annahme, dass ein Datensatz vorliegt in welchem die bekannten und zu erlernenden Muster mit einer Zielvariablen markiert sind. Der Datensatz enthält mehrere Attribute und eine Zielvariable. Die Verfahren lernen auf Grund der Ausprägungen der Attribute und die vorgegebene Zuordnung zu Klassen (Markierung)[334] die zugrundeliegende Logik zwischen den Attributen und der Zielvariablen. Wenn in Zukunft Datensätze mit den Attributen ohne der Zielvariablen vorliegen, so kann mit der erlernten Logik eine automatische Markierung oder Klassifikation vorgenommen werden. Neben der Annahme, dass ein markierter Datensatz vorliegt, wird angenommen, dass sich zukünftige Daten ähnlich verhalten damit die Klassifikation Gültigkeit hat. Klassifikationsverfahren sind typische Vertreter

[330] Vgl. Handl, A. (2010), S. 3.
[331] Vgl. Backhaus, K. et al. (2016), S. 15.
[332] Es werden überwachte Lernverfahren auf Datensätze angewandt, die markierte und unmarkierte Datensätze enthalten. (Vgl. Witten, I. H.; Frank, E. (2005), S. 294.
[333] Auch reinforcment learning. Autonome Systeme (Agenten) agieren mit deren Umgebung und sollen lernen die richtigen Entscheidungen zu treffen, um einen Zielzustand zu erreichen oder ein gewünschtes Verhalten zu erlernen. Die einzige Vorgabe von außen ist ein Belohnungssystem/Bestrafungssystem für richtige bzw. falsche Entscheidungen. (Vgl. Mitchell, T. M. (1997), S. 367.
[334] Aus als Label oder Target bekannt.

der Methoden des überwachten Lernens.[335] Statistisch sind sie mit den strukturprüfenden Verfahren vergleichbar.

Unüberwachte Lernverfahren funktionieren ohne einen vorab markierten Datensatz. Die Muster sind unbekannt und müssen gefunden werden. Die Clusteranalyse oder die Assoziationsanalyse sind Methoden des unüberwachten Lernens.[336] Statistisch sind die Verfahren mit der Strukturentdeckung vergleichbar.

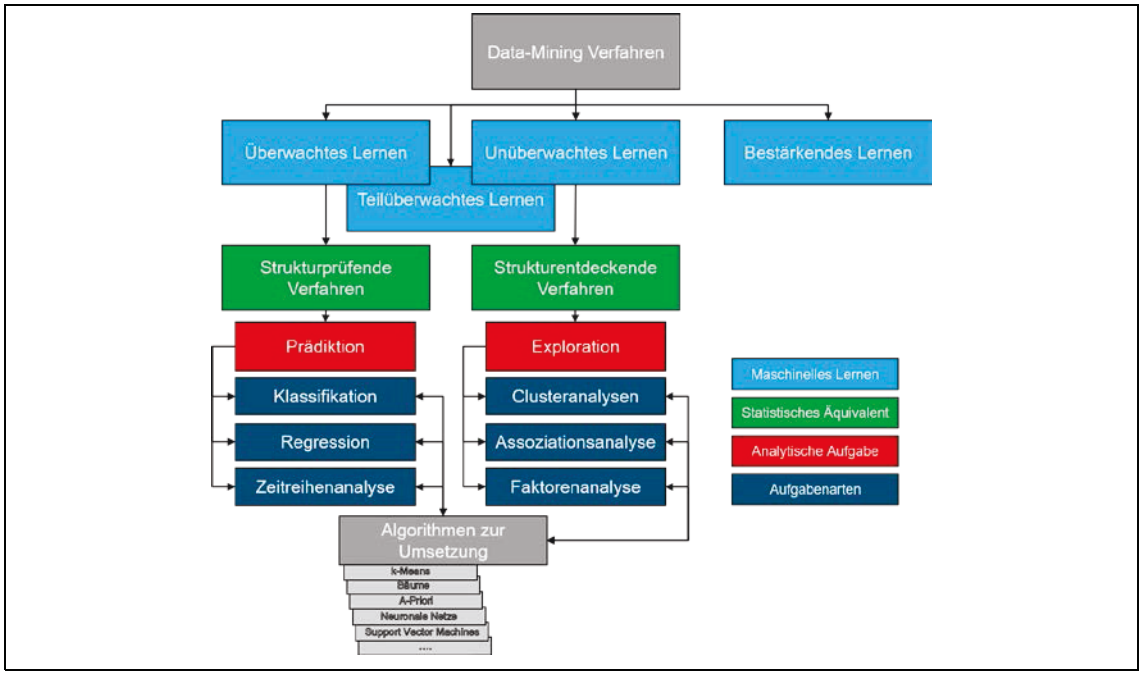

Abbildung 22: Gliederung von Data-Mining Verfahren[337]

Eine allgemeingültige Einteilung der Methoden gestaltet sich schwierig. BEEKMANN nimmt die Zuordnung der Verfahren zu den Aufgabenstellungen vor. Dabei zeigt sich das grundlegende Problem, dass einzelne Verfahren, wie Entscheidungsbäume und künstliche neuronale Netze für unterschiedliche Aufgabenstellungen verwendet werden können.[338] Seine Einteilung kann generisch auf die Prädiktion und die Exploration heruntergebrochen werden. Abbildung 22 gibt einen Überblick über die Gliederungsmöglichkeiten der Data-Mining Verfahren.

Data-Mining Verfahren

Die Data-Mining Verfahren werden kurz erörtert, um besser einzuschätzen, für welchen Fragestellungen sie sich als Lösung von Problemen eignen.

Klassifikationsverfahren:

Klassifikationsverfahren bestimmen ob ein Datensatz, bestehend aus einer Reihe von Attributen, zu einer vorgegeben Klasse gehört. Die Zuteilung erfolgt nach einem vorab erlernten Modell. Um das Modell zu erlernen werden historische Daten verwendet, bei

[335] Vgl. Witten, I. H. et al. (2011), S. 40.; Cleve, J.; Lämmel, U. (2014), S. 55.
[336] Vgl. Abbott, D. (2014), S. 5.; Cleve, J.; Lämmel, U. (2014), S. 55.
[337] Quelle: Eigene Darstellung in Anlehnung an Beekmann, F. (2003), S. 31 f.; Witten, I. H. et al. (2011), S. 40; Abbott, D. (2014), S. 5.; Hall, P. et al. (2014), S. 5.; Gröger, C. (2015), S. 3.
[338] Vgl. Beekmann, F. (2003), S. 31 f.

denen die Zuordnung zu den Klassen bereits vorgegeben ist. Die Art der benötigten Daten hängt vom angewandten Algorithmus ab, bzw. die vorhandenen Daten bestimmen den Algorithmus, der eingesetzt werden soll. Das Skalenniveau der Attribute kann metrisch oder nicht-metrisch, die Klassen müssen in den meisten Fällen nominal skaliert sein.[339]

Regressionsverfahren:

Regressionsverfahren wenden das gleiche Lernprinzip wie Klassifikationsverfahren an. Die Daten müssen jedoch metrisch skaliert sein. Die logistische Regression bildet eine Ausnahme bei der Skalierung der Daten, die die Klassen beschreiben. Diese müssen binär skaliert sein.[340]

Zeitreihenanalyse:

Zeitreihenanalysen sind Prädiktionen von metrisch skalierten Daten ohne eine Zielklasse. Während bei der Regression mehrere unabhängige Variablen verwendet werden, um eine abhängige Zielvariable vorherzusagen, erfolgt hier die Vorhersage einer Variablen rein aufgrund ihrer historischen Werte. Die Daten müssen metrisch skaliert sein.[341]

Anomalieerkennung:

Die Anomalieerkennung identifiziert Datensätze, die sich außerhalb der Norm des Gesamtdatenbestandes verhalten. Sie wird auch als Ausreißeranalyse verwendet. Sie ist klassisch ein unüberwachtes Verfahren, da Anomalien in der Regel unbekannt sind. Gibt man bereits bekannte Anomalien vor werden überwachte oder teilüberwachte Lernverfahren angewandt.[342]

Clusteranalyse:

Sie identifiziert Gruppen in Daten. Dabei werden die Ähnlichkeiten oder Unähnlichkeiten der Datensätze bezüglich ihrer Attribute bestimmt. Die Datensätze in einer Gruppe sollen möglichst homogen sein, während die Gruppen untereinander möglichst heterogen sein sollen.[343]

Faktorenanalyse:

Die Faktorenanalyse hat zum Ziel die Variablenmenge zu reduzieren. Das ist speziell bei Big Data Anwendungen hilfreich, um die Datenmenge für weitere Berechnungen zu verringern. Es wird dabei die Frage beantwortet, ob sich die gemessenen Variablen auf wenige reduzieren lassen, bzw. ob diese gebündelt werden können. Dabei wird unterstellt, dass sich bei vielen Variablen einige überlappen und das System überbestimmen.[344]

[339] Vgl. Kotu, V.; Deshpande, B. (2015), S. 11 & 408 ff.
[340] Vgl. Kotu, V.; Deshpande, B. (2015), S. 11 & 411.
[341] Vgl. Kotu, V.; Deshpande, B. (2015), S. 305.
[342] Vgl. Chandola, V. et al. (2009), S. 1 ff.
[343] Vgl. Pascual, D. G. (2015), S. 217.
[344] Vgl. Backhaus, K. et al. (2016), S. 386.

Assoziationsanalysen:

Die Assoziationsanalyse findet Beziehungen zwischen Attributen und Attributausprägungen von Datensätzen bzw. Transaktionen. Die Daten müssen für die Anwendung nominal skaliert sein. Metrische Daten können entsprechend transformiert werden.[345] Die Assoziationsanalyse eignet sich für eine datengestützte Schwachstellenanalyse gut. Sie fördert Regeln zu Tage die weiter als kausale Ursache-Wirkungszusammenhänge interpretiert werden können.

Datenanalysekonzepte

Die Analysekonzepte lassen sich auch nach der Fragestellung unterscheiden, die sie beantworten. Abbildung 23 zeigt die Gliederung der Analysekonzepte nach Komplexität der Analysen und des Wertes für das Unternehmen. Dabei kann auch die Grenze zwischen BI und Industrial Analytics oder Big Data Analytik gezogen werden. Die Diagnostic-Analytics bildet den Übergangsbereich zwischen den beiden Metabegriffen.[346] Je nach Komplexität der Diagnose ist es mit OLAP Verfahren möglich die Ursachen zu finden und folglich die Zuordnung zu BI oder es sind ausgeklügelte Data-Mining Verfahren nötig, was die Zuordnung zu Industrial Analytics rechtfertigt.

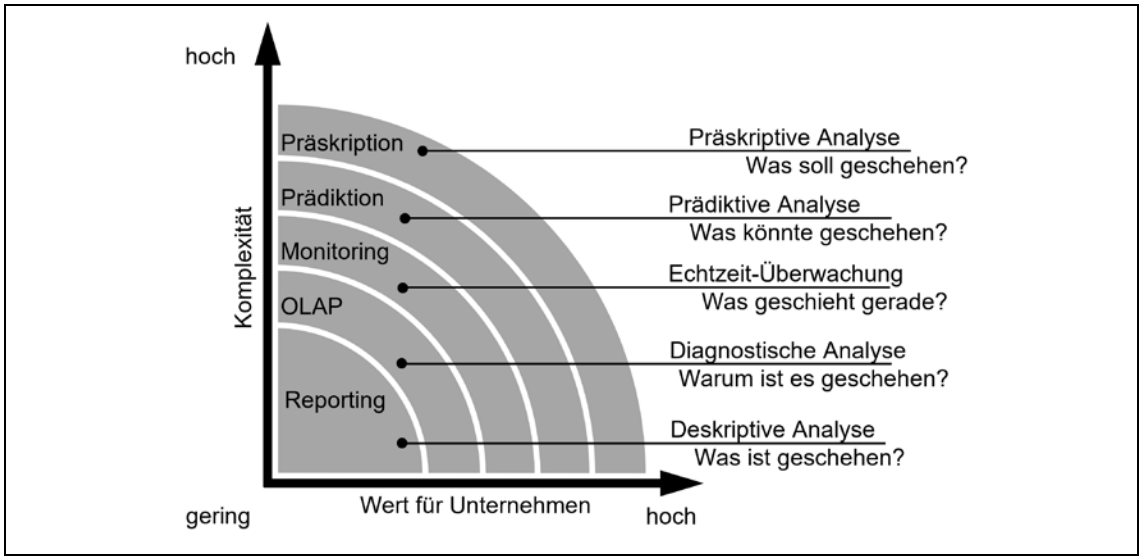

Abbildung 23: Arten der Analysekonzepte[347]

Die einfach deskriptive Analyse ist vergangenheitsorientiert und beantwortet die Frage was geschehen ist, ohne eine weitere Wertung vorzunehmen. Die diagnostische Analyse stellt sich die Frage warum etwas geschehen ist. Es sollten Ursachen und Korrelationen für Vorfälle gefunden werden. In manchen Fällen wird die diagnostische Analyse der deskriptiven zugeordnet. Das Monitoring stellt Ansprüche an die Datenerfassung und Übertragung der Daten. Es ist stark mit der deskriptiven Analyse verbunden, da Echtzeitdaten schnell visualisiert werden sollen. Die prädiktive Analyse ist mit der Frage konfrontiert Zukünftiges vorherzusagen. Dazu bedarf es eines historischen Datenbestandes und einer gut definierten Zielgröße mit ausreichend

[345] Vgl. Abbott, D. (2014), S. 145.
[346] Vgl. Ereth, J.; Kemper, H.-G. (2016), S. 459 f.
[347] Quelle: Dorschel, J. et al. (2015), S. 56.

beschreibenden Variablen, um prognostizierbare Muster zu erkennen. Die präskriptive Analyse gibt vor was geschehen soll. Basierend auf den Prognosen und den Analysen über die Ursachen, kann mit Zielfunktionen eine ideale Handlung vorgeschlagen werden. Die Anforderungen an ein System für solche Analysen sind besonders hoch, der Wert für das Unternehmen jedoch auch.[348]

4.2.2 Assoziationsanalyse

Die Assoziationsanalyse wird in der Literatur auch unter den Bezeichnungen *Assoziationsregeln*, *item set mining*, *frequent item set mining* oder *Warenkorbanalyse* geführt. Die Aufgabe der Analyse ist das Finden von Zusammenhängen zwischen Variablen. Die Herausforderung liegt in der effizienten Implementierung der Algorithmen, die möglichst schnell die verschiedenen Kombinationen des Zusammenspiels iterieren können. Uninteressante Stränge müssen frühzeitig gefunden und vernachlässigt werden, um in großen Datenmengen mit hunderttausenden Zeilen und duzenden Spalten die Rechenzeit minimal zu halten. Die bekanntesten Algorithmen sind dabei Apriori[349] und FP-growth.[350]

Allgemeines

Die Datenbasis ist in Form von Datensätzen organisiert wobei ein Datensatz als Transaktion bezeichnet wird. Eine Transaktion wird zur gleichen Zeit, bzw. im gleichen Zeitraum durchgeführt. Die Definitionen des Zeitstempels und des Begriffs Gleichzeitigkeit ist bei dieser Analyse essentiell. Eine Transaktion besteht aus einer Anzahl von Items. Items der gleichen Art sind Attribute im Datensatz. Werden mehrere Items zusammengefasst werden sie als Itemmenge bezeichnet.[351] Abbildung 24 fasst die Darstellung und Bezeichnung der Daten in der Assoziationsanalyse zusammen.

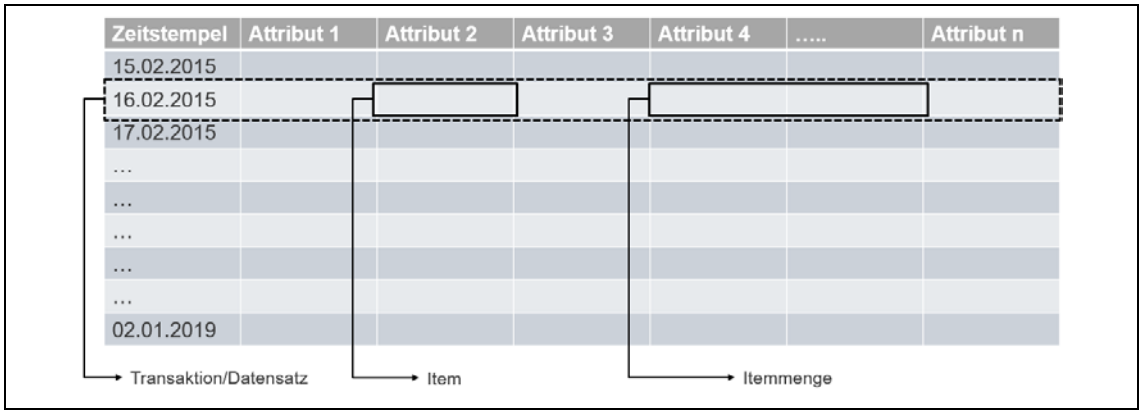

Abbildung 24: Datenstruktur Assoziationsanalyse[352]

[348] Vgl. Dorschel, J. et al. (2015), S. 56 f.
[349] Siehe Bollinger, T. (1996), S. 258 f.
[350] Vgl. Brin, S. et al. (1997), S. 255 ff.
[351] Vgl. Hettich, S.; Hippner, H. (2001), S. 427.
[352] Quelle: Eigene Darstellung

Die Assoziationsanalyse verwendet nur nominal skalierte Daten. Metrische Skalen müssen überführt werden. Daher arbeitet die Analyse nur mit Wahrscheinlichkeiten und Häufigkeiten.

Der Zusammenhang zwischen Items oder Itemmengen wird als Regel ausgegeben. Eine Regel besteht aus einem Regelrumpf, auch Prämisse, und einem Regelkopf, auch Konklusion. Sie ist so zu lesen: Wenn die Prämisse eintritt, dann führt dies zur Konklusion.[353]

Sind zwei Itemmengen[354] A und B Teil diese Regel und wenn A eintritt führt das zu B, so wird diese wie in Beispiel 4-4 folgt dargestellt:

Allgemeine Regelschreibweisen	$Regelrumpf \rightarrow Regelkopf$	4-1
	$Prämisse \rightarrow Konklusion$	4-2
Langschreibweise	$Itemmenge\ A \rightarrow Itemmenge\ B$	4-3
Kurzdarstellung	$A \rightarrow B$	4-4

Wichtig ist, dass die Itemmengen disjunkt sind. Umfasst die Itemmenge A die Items a und b, so darf die Itemmenge B diese beiden nicht enthalten.[355]

Interessantheitsmaße

Eine Regel wird mit Interessantheitsmaßen bewertet, die Auskunft über die Wichtigkeit und Aussagekraft der Regel geben. Die wichtigsten Regeln sind der Support, die Konfidenz und der Lift.[356] Folgend werden die drei Maßeinheiten kurz beschrieben.

Support:

Dieser ist die Wahrscheinlichkeit bzw. die relative Häufigkeit, dass zwei Itemmengen in der Gesamtmenge *D* der Transaktionen *t* gleichzeitig eintreten. Der Support der Regel $A \rightarrow B$ berechnet sich wie folgt:[357]

$$support(A \rightarrow B) = \frac{|\{t \in D | (A \cup B) \subseteq t\}|}{|D|} \quad \text{4-5}$$

Der Support dient als Filterkriterium, welches die Performanz der Algorithmen zur Regelableitung verbessern kann. Des Weiteren kann bei der Regelauswahl für die weitere Interpretation der Fokus nur auf Regeln mit einem hohen Support gelegt werden, da diese häufiger vorkommen und relevanter sind.

Konfidenz:

Eine Regel $A \rightarrow B$ muss nicht immer gelten d.h. wenn die Itemmenge *A* auftritt muss nicht immer die Itemmenge *B* eintreten. Die Konfidenz versucht speziell dieses Szenario zu beurteilen, wie oft die Itemmenge *B* tatsächlich eintritt, wenn die Itemmenge *A* vorkommt. Formel 4-6 zeigt die Definition der Konfidenz. Der zweite Teil über den

[353] Vgl. Hettich, S.; Hippner, H. (2001), S. 427.
[354] Es wird nur von Itemmengen gesprochen, da ein Item eine Menge mit der Anzahl 1 ist.
[355] Vgl. Bollinger, T. (1996), S. 258.
[356] Vgl. Beekmann, F. (2003), S. 85 ff.; Abbott, D. (2014), S. 149 f.
[357] Vgl. Bollinger, T. (1996), S. 258.

Support, zeigt die alternative Betrachtungsweise der Konfidenz. Nachdem der Support eine Wahrscheinlichkeit ist, ist die Konfidenz die bedingte Wahrscheinlichkeit, dass B unter der Bedingung von A eintritt.[358]

$$confidence(A \rightarrow B) = \frac{|\{t \in D | (A \cup B) \subseteq t\}|}{|\{t \in D | A \subseteq t\}|} = \frac{support(A \rightarrow B)}{support(A)} \qquad 4\text{-}6$$

Während der Support symmetrisch ist – $support(A \rightarrow B) = support(B \rightarrow A)$ – gilt das für die Konfidenz nicht. Es ist so die wahrscheinlichere Wirkungsrichtung – Ursache-Wirkungsbeziehung – feststellbar.

Lift:[359]

Die Berechnung des Lifts beseitigt Schwächen, die in der Konfidenz zu finden sind. Diese vernachlässigt die Wahrscheinlichkeit von der Itemmenge in der Prämisse. Somit ist es möglich, dass zwei statistisch unabhängige Itemmengen eine hohe Konfidenz haben. Der Umstand ergibt sich daraus, dass die Konfidenz Veränderungen der Grundgesamtheit D oder der Menge der Itemmenge in der Prämisse außer Acht lässt. Der Lift wirkt dem entgegen, indem die Wahrscheinlichkeit der Itemmenge der Prämisse in der Berechnung gesondert berücksichtigt wird.

Formel 4-7 zeigt die Berechnung des Lifts. Die Konfidenz wird in Beziehung zur Wahrscheinlichkeit, dass Itemmenge B in der Gesamtmenge vorkommt, gesetzt. Sind Zähler und Nenner gleich groß, nimmt der Lift den Wert 1 an und die beiden Itemmengen der Regel sind statistisch unabhängig und nicht korreliert. Die Konfidenz besagt, dass B eintritt falls A eintritt, wenn B jedoch gleich wahrscheinlich in der Gesamtmenge vorkommt, ist der Zusammenhang der Konfidenz Zufall, speziell bei einer Konfidenz von 1, die besagt, dass die Regel immer gilt.

$$lift(A \rightarrow B) = \frac{confidence(A \rightarrow B)}{support\ (B)} \qquad 4\text{-}7$$

Der Lift besagt hingegen, um wieviel wahrscheinlicher die Itemmenge B in der Menge von Transaktionen vorkommt in der Itemmenge A beinhaltet ist, verglichen zur Wahrscheinlichkeit des Vorkommens in der Gesamtmenge. Der Wert 5 für den Lift der Regel $A \rightarrow B$ besagt somit, dass die Itemmenge B in den Transaktionen, in denen die Itemmenge A enthalten ist, fünfmal häufiger vorkommt als in der Gesamtmenge der Transaktionen. Die beiden Itemmengen sind miteinander positiv korreliert. Diese Regeln sind für eine weitere Betrachtung interessant. Ein Lift kleiner 1 bedeutet eine negative Korrelation. Der Regelzusammenhang ist somit unwahrscheinlicher. Diese Regeln sind nicht interessant. Der Lift ist wie der Support symmetrisch.

Selbst wenn die Interessantheitsmaße Schwächen der andern ausgleichen, sollte nicht nur eines für die Regelauswahl verwendet werden, da alle Regeln ihre Berechtigung haben. Die drei Maße sollten für alle Regeln angezeigt und darauf basierend eine Entscheidung getroffen werden.

[358] Vgl. Bollinger, T. (1996), S. 258.; Hettich, S.; Hippner, H. (2001), S. 477.
[359] Vgl. Hettich, S.; Hippner, H. (2001), S. 447.; Beekmann, F. (2003), S. 85 f.

4.2.3 Skalenniveaus

Analysealgorithmen verlangen bestimmte Skalenniveaus damit sie funktionieren. Das Skalenniveau bestimmt wie die Daten dargestellt werden. Es werden vier Skalen unterschieden:

- Nominalskala
- Ordinalskala
- Intervallskala
- Ratioskala

Das Skalenniveau bestimmt den Informationsgehalt der Daten und die Rechenoperationen, die mit ihnen durchgeführt werden können.

Die Nominal- und Ordinalskala sind nicht-metrische (kategoriale) Skalen. Sie haben den geringsten Informationsgehalt, wobei jener der Nominalskala geringer als jener der Ordinalskala ist. Die Nominalskala bildet Klassifizierungen ab, wie das Geschlecht, Farben, Automarken oder Fehlercodes. Es können Häufigkeiten bestimmt und diese ins Verhältnis gesetzt werden. Eine Sonderform der Nominalskala ist die Binärskala, die aussagt ob eine Merkmalsausprägung vorhanden ist (1) oder nicht (0)[360]. Die Ordinalskala erlaubt eine Rangordnung darzustellen. Ob Anlage A besser funktioniert als Anlage B und diese besser funktioniert als Anlage C. Der Abstand zwischen den Rängen ist nicht definiert. Es kann gesagt werden was besser ist, jedoch nicht um wieviel. Arithmetische Rechenoperationen sind daher unzulässig.

Die Intervallskala und die Ratioskala sind metrische (kardinale) Skalen. Mit ihnen können arithmetische Rechenoperationen durchgeführt werden. Die Intervallskala ist in gleich große Skalenabschnitte unterteilt. Es ist eine ordinale Reihung möglich und der Abstand zwischen den Klassen ist definiert. Ratingskalen[361] werden oft als Intervallskalen gehandhabt, sind jedoch ohne eine Bestätigung der Fixabstände zwischen den Ratingausprägungnen eigentlich Ordinalskalen. Die Celsiusskala zählt z. B. zu den Intervallskalen. Die Ratioskala enthält den höchsten Informationsgehalt. Sie besitzt im Gegensatz zur Intervallskala einen natürlichen Nullpunkt, der dem Nicht-vorhanden-sein entspricht. Die physikalischen Standardgrößen nach SI entsprechen dieser Definition, sowie ökonomische Größen wie Kosten oder der Preis.

Höher skalierte Daten sind vorzuziehen, da es immer möglich ist, diese in niederskalierte Skalen zu überführen. Die metrisch skalieren Daten können durch die Bildung von Klassen, wie in einem Histogramm, in nominal skalierte Daten überführt werden. Dieser Vorgang ist jedoch mit einem Informationsverlust verbunden. Umgekehrt, niedrige Skalenniveaus in höhere Skalenniveaus zu überführen, ist nicht möglich.[362]

[360] Vgl. Backhaus, K. et al. (2016), S. 363.
[361] Eine Zuordnung von z. B. 1 bis 10 vornehmen.
[362] Vgl. Backhaus, K. et al. (2016), S. 10 ff.

4.3 Kritische Würdigung und Relevanz für die Arbeit

Reifegradmodelle führen ihre Bewertung in erster Linie anhand von Prozessen durch und beurteilen auf diese Art die Reife von Unternehmen und Organisationen diese Prozesse durchzuführen. Für die Entwicklung eines Reifegradmodells zur Bewertung von Daten bzw. des datenanalytischen Prozesses ist es daher nötig einen Standardprozess zu definieren, der bewertet werden soll. Für diese Arbeit wird der CRISP-DM Prozess als Grundlage verwendet.

Der CRISP-DM Prozess hat gegenüber den anderen vorgestellten Modellen Vorteile, die zu seiner Auswahl führten. Er ist weit verbreitet und sehr gut dokumentiert. Das zyklische Vorgehen und die vorgelagerte Definition der Ziele macht ihn neben der generischen Beschreibung der Prozessschritte universell für Analysezwecke einsetzbar. In seiner Grundkonstruktion ist er ein laufender zyklischer Prozess, der die durchgeführten Schritte und Ergebnisse dokumentiert und reflektiert. Abstrakt betrachtet, bildet er einen KVP oder einen Qualitätsregelkreis ab. Beide Ansätze sind für die Übertragung auf ein Reifegradmodell wichtig. Das SEMMA Modell hat keine offensichtlich geplanten Iterationsschritte und kritisch betrachtet greift der KDD relativ spät im Prozessablauf ein, falls sich das gewünschte Ergebnis nicht ergibt. Der KDID hebt Subschritte des CRISP-DM auf die generische Ebene und verändert leicht die Folge der Prozessschritte des CRISP-DM, wie anhand der Erstellung der Vorstudie zu sehen ist. Kritisch betrachtet handelt es sich hier um die Erstellung des Testdesigns, welches im CRISP-DM in der Modellierung vorkommt. Da der CRISP-DM in seiner obersten Ebene generischer ist und weiterer verbreitet, wird ihm gegenüber des KDID der Vorzug gegeben.

Die ersten beiden Phasen dienen der Erhebung einer IST-Situation, um mögliche Hindernisse im Projektablauf antizipieren zu können. Diese IST-Erhebung ist ein wichtiger Bestandteil einer Reifegradbewertung. Das entwickelte Reifegradmodell soll hier ansetzen und das Vorgehen und die Ergebnisse dieser beiden Schritte konkretisieren. Die Durchführung der Analyse ist mit zwei Schritten im Kern abgebildet und beschreibt den Datenanalyseprozess im wesentlichen Umfang. Die letzten beiden Phasen befassen sich mit der Implementierung der Analyseergebnisse womit der CRISP-DM ein Datenanalyseprojekt ganzheitlich beschreibt.

Für die Anwendung im Reifegradmodell wird der CRISP-DM in drei Metaphasen unterteilt. In der Vorbereitungsphase wird das Reifegradmodell eingesetzt und bewertet den Prozess der Datenanalyse anhand der zwei weiteren Metaphasen. Die Durchführung beinhaltet die Datenanalyse und die Operationalisierung umfasst die Schritte zum Einsatz des erzeugten Modells. Die Vorbereitung wird bewusst nicht bewertet, da die relevanten Schritte eines datenanalytischen Prozesses bzw. Projektes in der Durchführung und Operationalisierung liegen. Neben der Einteilung der CRISP-DM Phasen in drei Metaphasen, wurden die Prozessschritte nochmals zusammengefasst, um die Verbindung mit Reifegradkategorien zu erleichtern. Tabelle 11 zeigt die Zusammenfassung in einer Übersicht, wobei die grau hinterlegten Spalten die neuen Einteilungen sind.

Ein wichtiger Einfluss in der Umsetzung datenanalytischer Aufgabenstellungen sind die einsetzbaren Methoden und Algorithmen. Die Algorithmen der Methoden haben in vielen Fällen spezifische Anforderungen an die Daten, die in einer Reifegraderhebung

berücksichtigt werden müssen. Sei es der Umfang der Daten für das Erlernen von Mustern oder ein wesentlicher Einflussfaktor in diesem Zusammenhang das Skalenniveau der Daten. Diese Faktoren müssen im Reifegradmodell berücksichtigt werden.

Die Assoziationsanalyse wurde als Algorithmus im Detail behandelt, da sie sich eignet, um Zusammenhänge in Daten und Wirkungsrichtungen zu finden. Sie bildet datenanalytisch dadurch eine Schwachstellenanalyse ab, anhand derer das Reifegradmodell entwickelt und getestet werden soll. Im Gegensatz zu anderen Analyseverfahren sind die Anforderungen an die Daten nicht so groß. Diese Rahmenbedingungen sind für die Testphase wichtig, da Analyse auch bei Unternehmen durchgeführt werden sollen, deren Datenqualität noch nicht auf dem höchsten Niveau ist. Da das Reifegradmodell anhand einer Schwachstellenanalyse entwickelt wird und der CRISP-DM Zyklus und die Assoziationsanalyse in diesem Zusammenhang zum Einsatz kommen, wird im folgenden Kapitel die Schwachstellenanalyse beschrieben, wobei der Fokus nicht rein auf die dieselbe gelegt wird.

Tabelle 11: Zusammenfassung der CRISP-DM Schritte[363]

Meta-phase		CRISP-DM Phase	Aggregierter Prozessschritt	CRISP-DM Prozessschritte
Vorbereitung		Geschäftsmodell verstehen und anpassen	Zielfestlegung und Bewertung	Unternehmensziele festlegen Data-Mining Ziele festlegen Umfeld bewerten
			Projektplan festlegen und Reifegradbestimmung	Projektplan erstellen
		Daten verstehen	Ausgangsdaten sammeln	Daten sammeln Daten beschreiben
			Datenqualität prüfen	Daten untersuchen Datenqualität prüfen
Datenanalyse	Durchführung	Daten aufbereiten	Vorbereitung	Daten auswählen Daten integrieren Daten formatieren
			Aufbereitung	Daten bereinigen Daten erstellen
		Modellierung	Testen	Modellierungsverfahren wählen Testdesign erstellen
			Finalisierung	Modell erstellen Modell bewerten
	Operationalisierung	Evaluation	Organisatorisch	Prozess bewerten Nächste Schritte planen
			Evaluierung	Ergebnisse evaluieren
		Einsatz	Organisatorisch	Endbericht erstellen Projekt prüfen und bewerten
			Planung	Einsatz planen Überwachung planen

[363] Quelle: Eigene Darstellung

5 Schwachstellenanalyse

Die Schwachstellenanalyse gewinnt aus mehreren Gründen zunehmend an Bedeutung. Zum einen steigt der Druck neue Potenziale zur Effizienzsteigerung zu finden und zum anderen nimmt die Komplexität von Anlagen und Prozesses laufend zu. Die fortschreitende Digitalisierung erlaubt in diesem Zusammenhang eine erleichterte Sammlung von Daten mit Hilfe derer komplexe Analysen aus dem Gebiet des Data-Minings durchgeführt werden können.

Folgend wird die Schwachstellenanalyse in den inhaltlich wichtigen Details erörtert. In einem modernen Asset Management ist sie ein wesentlicher Bestandteil zur Effektivitäts- und Effizienzsteigerung. Es werden die verschiedenen Klassifizierungsansätze für Schwachstellen und deren Identifikation mit Analysen betrachtet. In dieser Arbeit wird der Fokus auf die Bedeutung der Schwachstellenanalyse für die Instandhaltung gelegt. Um den Entwicklungen der Digitalisierung gerecht zu werden, wird am Ende eine Methode vorgestellt, die Data-Mining zur Identifikation von Ursachen von Schwachstellen basierend auf großen Datenmengen ermöglicht.

5.1 Schwachstellen und Ursachen

Laut DIN 31051 ist eine Schwachstelle eine „Einheit, bei der ein Ausfall häufiger als es der geforderten Verfügbarkeit entspricht, eintritt und bei der eine Verbesserung möglich und wirtschaftlich vertretbar ist.“[364]

Eine Schwachstelle tritt somit an einem Ort auf, ausgelöst durch eine Ursache und hat Einfluss auf einen Prozess in unterschiedlichen Verlustdimensionen. Wobei der Ort in der Regel eine physische Stelle ist, an der Daten erhoben werden können. Anhand der Daten kann die Schwachstelle identifiziert werden. Die Verlustdimensionen sind dabei weit gefächert und umfassen Zeit-, Mengen-, Qualitäts-, und Energieverluste.[365]

Die Unterscheidung der Schwachstellen ist in der Regel vielfältiger als die singuläre Fokussierung auf Maschinen und Anlagen wie in der Definition der Instandhaltungsnorm. Darin zeigt sich die Schwäche der Definition nach DIN 31051. Die Anlage wird als Schwachstelle betrachtet, da in produzierenden Unternehmen an ihr die Zielerfüllung gemessen wird. Die Definition nach DIN 31051:2012 ist bereits eine Verbesserung gegenüber der älteren Definition, in der die Schwachstelle als eine Schadstelle an einer Anlage betrachtet wurde. In den älteren Ausgaben der DIN 31051 wurde der Begriff Schaden definiert als:

„Veränderung an einem Bauteil, durch die seine vorhergesehene Funktion beeinträchtigt oder unmöglich gemacht wird oder eine Beeinträchtigung erwarten lässt.“[366] Diese sehr eingeschränkte Sicht auf die Schwachstelle, als rein technische Schadensstelle, wird in

[364] Deutsches Institut für Normung (2012), S. 7.
[365] Vgl. Henning, J. (2009), S. 153.
[366] Strunz, M. (2012), S. 96.

dieser Arbeit nicht geteilt. Die Schwachstelle kann an unterschiedlichen Orten auftreten und ist nicht rein durch einen Totalausfall aufgrund eines Schadens definiert.

Sie steht meist am Ende einer Kette von Ereignissen und Einflüssen, die zur Verringerung der Verfügbarkeit führte. Die Kette von Ereignissen und Einflüssen werden als die Ursachen der Schwachstelle gesehen. Kritisch betrachtet ist die Schwachstelle nur die Wirkung der Abfolge und Verkettung unterschiedlicher Ursachen. MEXIS definiert daher Schwachstellen allgemeiner:

„Einer Schwachstelle ist der kausale Ereignisort der Ursachenauswirkung in technischen Konstruktionen."[367]

Die Begriffe Schwachstelle und Ursache werden oft synonym verwendet. Zum Beispiel HENNING spricht von:[368]

- Maschinentechnischen Schwachstellen
  - Analgen und deren Komponenten sind die Ursachen dafür.
- Verarbeitungstechnischen Schwachstellen
  - Die verarbeiteten Stoffe und Produkte sind die Ursachen hier.
- Organisatorischen Schwachstellen
  - Die Organisation ist für die Verluste verantwortlich.
- Sonstige Schwachstellen:
  - Alle sonstigen Verlustquellen, wie z.B. Bedienfehler.

Henning identifiziert neben diesen vier Faktoren auch die Verkettung der Anlagen im Produktionsprozess als einen möglichen Einfluss auf die Produktivität. Der Unterschied zwischen Reihen- und Parallelverkettung oder eine Kombination daraus, erhöht die Komplexität der Schwachstelle.[369]

Es zeigen sich Wirkungen in der Durchführung eines Prozesses, die unterschiedliche Ursachen haben können. Davon sind Störungen bei Anlagenelementen nur ein Teil davon. Diese sind jedoch nicht begrenzt zu betrachten, da sich defekte Anlagenelemente, speziell bei verketteten Anlagen gegenseitig beeinflussen können was zu Folgeausfällen führen kann. Häufige Rüstvorgänge, durch geringe Losgrößen wegen eines kundenindividuellen Produktionsprogramms sind weitere Ursachen für die Verringerung der Produktivität und der Verfügbarkeit einer Anlage und eines Prozesses.[370] Eine starke Abgrenzung in die vier oben genannten Schwachstellenarten mit den Ursachen ist daher nicht möglich, da es unterschiedliche Einflüsse gibt.

BIEDERMANN fasst Schwachstellen als „Zielabweichungen und Mängel im Planungsorganisations- und Informationsdesign bzw. […] Fehler und Mängel eines Bauteils oder Systems" auf.[371]

Die Schwachstelle wird mit dieser Definition vom reinen physischen Objekt gelöst. Die Betrachtungseinheit umfasst dadurch Prozesse und organisatorische Gegebenheiten. Schwachstellen werden in fünf Klassen unterteilt. Jene Schwachstellen, die sich durch

[367] Mexis, N. D. (1992), S. 96.
[368] Vgl. Henning, J. (2009), S. 155.
[369] Vgl. Henning, J. (2009), S. 158.
[370] Vgl. Henning, J. (2009), S. 160 f.
[371] Biedermann, H. (2017), S. 28.

unzureichende Infrastruktur und informationstechnische Unterstützung zeigen. Darüber hinaus Schwachstellen, die sich durch Unzulänglichkeiten in der Ablauforganisation zeigen. Schwachstellen in der Aufbauorganisation bilden die dritte Gruppe und komplettieren die prozessualen und organisatorischen Schwachstellen. Die objektbezogenen Schwachstellen als vierte Gruppe umfassen die klassischen Schwachstellen an Anlagenteilen. Die letzte Gruppe beinhaltet die Schwachstellen der Infrastruktur und der Umgebung der Anlage.[372]

5.2 Einsatz und Verfahren der Schwachstellenanalyse

Die Identifikation und Analyse von Schwachstellen geht durch die oben differenzierte Klassifikation von Schwachstellen über die reine Fehleranalyse hinaus, zu der sie in der DIN 31051 gezählt wird.[373] Die Schwachstellenanalyse selbst wird dabei nicht genauer definiert. Es wird differenziert, welche Arten der Schwachstellenanalyse es gibt und mit welchen Methoden die Schwachstellenermittlung durchgeführt wird. Um die Methoden einzusetzen sind unterschiedliche Daten nötig, die sich an dem Ziel der Analyse orientieren. Diese müssen in der Lage sein, Schwachstellen zu identifizieren, die über ein physisches Objekt wie ein Bauteil hinausgehen. Dazu zählen neben virtueller bzw. abstrakter Auftretensorte wie Organisationseinheiten auch organisatorische Gegebenheiten oder Prozessschritte.

5.2.1 Arten der Schwachstellenanalyse

Abhängig vom Ziel oder des Einsatzgebietes der Analyse können unterschiedliche Arten der Schwachstellenanalyse unterschieden werden. Die technische Schwachstellenanalyse untersucht den Einfluss technischer Parameter auf Funktionen. Instandhaltungstechnische Schwachstellenanalysen beschäftigen sich mit der Untersuchung von Abnutzungsvorräten von Bauteilen. Sie hat mit der Erweiterung der Aufgaben der Instandhaltung um die Verbesserung, zusätzlich an Bedeutung gewonnen. Organisatorische Schwachstellenanalysen untersuchen Zusammenhänge, die nicht direkt von einer physischen Betrachtungseinheit ausgelöst werden, sondern zur Funktion beitragen. Dazu zählen die Bedienung, das Personal oder die Planung. [374]

Für eine präzisere Identifikation der Schwachstellen und deren Ursachen wird die Unterteilung in die fünf Klassen nach BIEDERMANN verwendet. Analog dazu gestaltet sich die Gliederung der Arten der Schwachstellenanalyse. Die Analyse zur Feststellung von Schwachstellen in der technischen Infrastruktur findet Unzulänglichkeiten im Datenmanagement. Ablauforganisatorische Schwachstellenanalysen decken Probleme auf, die ihre Ursachen in der Prozessdurchführung haben. Dazu zählen einerseits ein hoher Formalisierungsgrad und eine geringe Standardisierung und andererseits nicht wertschöpfende Arbeitsschritte, die den Prozess ineffizient machen. Schwachstellenanalysen der Aufbauorganisation identifizieren Probleme und Ursachen,

[372] Vgl. Biedermann, H. (2017), S. 28 f.
[373] Vgl. Deutsches Institut für Normung (2012), S. 12.
[374] Vgl. Mexis, N. D. (1992), S. 97 f.; Vgl. Ryll, F.; Freund, C. (2010), S. 24.

die sich u. a. durch nicht eindeutige Entscheidungsbereiche oder schlechte Personalkapazitätsausstattungen ergeben. Die objektbezogene Schwachstellenanalyse identifiziert die Versagensarten und deren Ursachen und sie gibt darüber hinaus Auskunft über die Beeinflussung der Komponenten untereinander im Gesamtsystem. Die fünfte Art der Schwachstellenanalyse beschäftigt sich mit den Auswirkungen und Ursachen, die durch die Infrastruktur und die Umgebung induziert werden.[375]

MEXIS' gliedert die Herangehensweise zur Ermittlung von Schwachstellen in drei Kategorien. Die kenngrößenbezogene Schwachstellenermittlung ist jene die am weitesten verbreitet ist. Durch die Bildung und Aufschreibung von Kenngrößen rund um den Analgenausfall, werden Auswertungen ermöglicht, die auf die Schwachstelle hindeuten. Gut qualifizierte Mitarbeiter, die sich der Notwendigkeit bewusst sind, exaktes Datenmaterial aufzuschreiben, sind dafür eine Voraussetzung.[376]

Die Schadensstatistikmethode wertet Schäden kategorisiert nach Komponenten und Fehlerarten aus. Die so erfassten Daten werden mit Normwerten verglichen, um daraus Erkenntnisse über den Schadensverlauf und die Ursachen ziehen zu können. Die Datenerfassung erfolgt in standardisierten Zeitabständen und Schritten.[377]

Die Entwicklungen der Digitalisierung tragen dazu bei, die Aufzeichnungen für beide Verfahren zu verbessern. MDE- und BDE-Systeme erfassen Ausfälle teilweise automatisiert, bzw. den Mitarbeitern wird durch mobile Systeme und hinterlegte Vorgehensweisen bei der korrekten Datenerfassung assistiert.[378]

Der entscheidende Nachteil der beiden Verfahren besteht darin, dass die Daten durch Ausfälle erst anfallen müssen. Die kausalitätsbezogene Schwachstellenanalyse versucht Ausfälle vorab zu vermeiden, indem unterschiedliche Möglichkeiten, die zum Ausfall führen können, vorab erdacht und analysiert werden. Speziell in der Entwicklungsphase sollten die Verfahren ihren Einsatz finden. Wobei es keine Rolle spielen sollte, ob technische Systeme wie Anlagen oder ob Prozesse entwickelt werden.[379]

Daten für die Schwachstellenanalyse

Für eine Schwachstellenanalyse werden Daten benötigt, die direkt mit dem Ausfall in Verbindung gebracht werden können. Das sind Zeitdaten, die auf die Dauer eines Ausfalls hindeuten oder Konstruktionsdaten, die über die Anlagenstruktur und die betroffenen Komponenten Auskunft geben. Die Ursachendaten ermöglichen eine Aussage über Ort und Art der Ursache für eine Schwachstelle. Arbeitsdaten und Kostendaten geben Auskunft über die Art, Dauer und den Umfang von Arbeiten im Rahmen der Schwachstellenbeseitigung sowie die Kosten, die durch die Schwachstelle anfallen können.[380] Die Ausführungen von MEXIS und HASTINGS fokussieren sich stark auf Ausfälle von Anlagen. Sie lassen sich jedoch generalisieren, da grundsätzlich immer

[375] Vgl. Biedermann, H. (2017), S. 28 f.
[376] Vgl. Mexis, N. D. (1992), S. 98.
[377] Vgl. Mexis, N. D. (1992), S. 98 f.; Strunz, M. (2012), S. 158.
[378] Vgl. Biedermann, H. (2017), S. 31.
[379] Vgl. Mexis, N. D. (1992), S. 100.; Biedermann, H. (2017), S. 31.
[380] Vgl. Mexis, N. D. (1992), S. 153.; Hastings, N. A. J. (2015), S. 223 f.

jene Daten zur Analyse zur Verfügung stehen müssen, die in direktem Zusammenhang mit dem Auftreten der Schwachstelle stehen.

Um die Analyse. Auswertung und Einordnung der Daten zu erleichtern, empfiehlt sich ein standardisiertes Schema vorzugeben. Ein Teil davon ist die Zuordnung der Schwachstelle zum Ort des Auftretens. Dieser ist entweder physisch, wie ein Teil einer Anlage oder nicht physisch, wie eine Organisationseinheit oder ein Arbeits- oder Prozessschritt. Die weiteren gehen genauer auf Art und die Ursache der Schwachstelle ein. Wesentliche Informationen sind dabei das Schadensbild und die Schadensursache. Das Schadensbild beschreibt den Schaden, bzw. die Schwachstelle anhand äußerer Merkmale. Bei physischen Elementen sind Risse, Brüche oder Abtragungen gängige Beispiele. Bei abstrakten Elementen, wie Systemen, sind Leistungsdefizite oder Ungenauigkeiten mögliche Manifestationen von Schwachstellen. Die Schadensursache gibt Auskunft über den Grund des Schadens oder der Schwachstelle. Beispiele sind Überlasten oder fehlende Schmierung bei physischen Schäden. Organisatorische Ursachen sind u. a. im Prozessablauf zu finden, wie zum Beispiel häufiges Rüsten oder eine gestörte Materialzuführung.[381] Die Daten können in ein Codierungsschema überführt werden. In Kombination mit mobilen Erfassungsgeräten kann damit der Mitarbeiter durch den Meldungsprozess geführt werden. Dies erleichtert die Eingabe von Störungen. Infolge dessen erhöht sich die Datenqualität und das Ergebnis von Analysen wird verbessert.[382]

Diese Daten sind der Input für einfache Methoden der Schwachstellenanalyse. Die horizontale und vertikale Integration erlaubt es Daten aus unterschiedlichen Quellen zusammen zu führen und komplexe Big Data gestützte Analysen durchzuführen. Die Möglichkeiten der verwendeten Daten reichen bis zu Sensormessungen an einzelnen Anlagenkomponenten.[383]

Instrumente zur Schwachstellenanalyse

Ein oft eingesetztes Instrument zur Schwachstellenermittlung ist die ABC-Analyse. Sie strukturiert Daten nach dem Paretoprinzip. Das Prinzip beruht auf dem oft beobachteten Umstand, dass ein kleiner Teil möglicher Ursachen für einen Hauptanteil der Wirkungen verantwortlich ist.

Zielabweichungen sowie Kostenüberschreitungen sind ein häufiger Auslöser von Schwachstellenanalysen. Schwachstellen, die hohe Folgekosten verursachen, sind wirtschaftlich vertretbar - temporär oder permanent - zu beseitigen.[384]

Unter Kosten wird der mit Preisen bewertete Verzehr von Produktionsfaktoren, wie Güter oder Dienstleistungen, der durch die betriebliche Leistungserstellung und -verwertung verursacht wird, verstanden.[385] Die Instandhaltungskosten umfassen den Verzehr, der sich durch die Aufgaben der Instandhaltung ergibt. Dazu zählen die Tätigkeiten im Rahmen der Grundtätigkeiten der Instandhaltung aus Abbildung 26. Sie lassen sich in direkte und indirekte Instandhaltungskosten gliedern. Zu den Direkten zählen Personal,

[381] Vgl. Nebl, T.; Prüß, H. (2006), S. 183 ff.; Hastings, N. A. J. (2015), S. 225 f.
[382] Vgl. Pawellek, G. (2016), S. 122-130.
[383] Vgl. Biedermann, H. (2018), S. 27.
[384] Vgl. Pawellek, G. (2016), S. 141.; Bernerstätter, R.; Kühnast, R. (2017), S. 166.
[385] Vgl. Wöhe, G. et al. (2016), S. 290, 638.

Material und Ersatzteile sowie Energie oder Instandhaltungsleistungen von Fremdfirmen. Die indirekten Instandhaltungskosten umfassen die Ausfallkosten.[386]

Die Ausfallkosten im engeren Sinne sind jene Kosten die durch Erfolgseinbußen, Erlösminderungen und während der Durchführung von Instandhaltungsmaßnahmen anfallen. Sie stehen mit einem eintretenden Ausfall in Verbindung. Die Ausfallkosten im weiteren Sinne sind jene Kosten, die durch Maßnahmen anfallen die der Bekämpfung negativer ökonomischer Konsequenzen von Verschleißerscheinungen im Beschaffungs-, Produktions-, und Absatzbereich dienen.[387][388]

In der Instandhaltung werden die Instandhaltungskosten als erste Datenquelle für die ABC-Analyse verwendet. Die Anlagen oder Komponenten werden entsprechend der Höhe der Kosten aufsteigend sortiert und die Kosten kumuliert. Es ergibt sind ein Bild ähnlich jenem in Abbildung 25. Die Verteilung in Linie 1 kann in den meisten Fällen in die drei Klassen A, B, C mit deren groben Grenzen geteilt werden. Die Verteilung von Linie 2 würde sich ergeben, wenn die Kosten bei jeder Betrachtungseinheit gleich wären.[389]

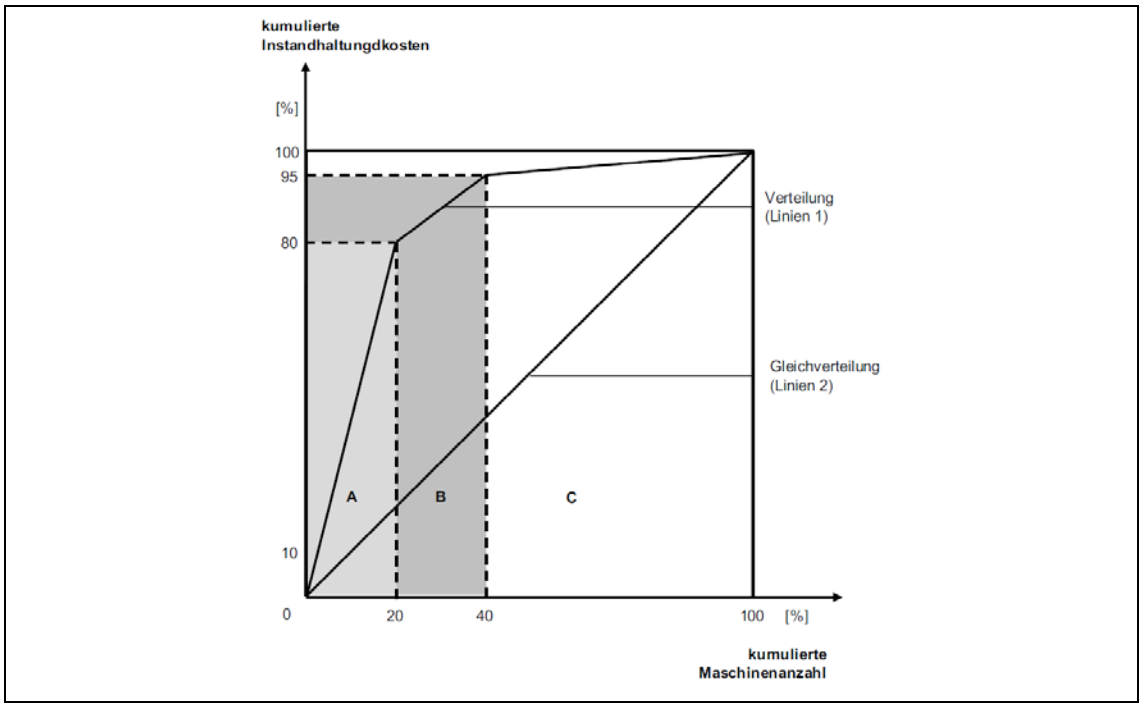

Abbildung 25: Typische Verteilung einer ABC-Analyse[390]

Die Betrachtungseinheiten der Klasse A würden als Schwachstellen klassifiziert werden und müssten genauer untersucht werden. Um zu vermeiden, dass ein reiner Kostenfokus in der Analyse gelegt wird, kann die gleiche Analyse mit anderen Kenngrößen durchgeführt werden. Die Betrachtungseinheiten der A-Klasse werden verglichen und jene genauer analysiert, die in allen ABC-Analysen der A-Klasse

[386] Vgl. Biedermann, H. (2008), S. 43.
[387] Vgl. Biedermann, H. (2008), S. 125.
[388] Vgl. Biedermann, H. (1985), S. 224.
[389] Vgl. Pawellek, G. (2016), S. 141.
[390] Quelle: Pawellek, G. (2016), S. 142.

zugeordnet werden. So kann verhindert werden, dass Ausreißer bei einer Kenngröße das Gesamtbild verzerren.[391]

Die verdichtete Auswertung von Kennzahlen und die Darstellung im zeitlichen Verlauf als Kennzahlenprofil ist ein Instrument, welches neben der Schwachstellenanalyse langfristig der Instandhaltungsstrategieanpassung dienen soll. Neben den objektbezogenen Kennzahlen wie MTTR, MTBF, Ersatzteilvorrat oder Lohnkostenanteil hat sich die Darstellung der Kostenanteile an den Gesamtinstandhaltungskosten bewährt. Durch die Aufgliederung in z. B. Wartungs-, Inspektions- und Instandsetzungskosten können u. a. Verbesserungspotenziale in der Instandhaltungsstrategie identifiziert werden. Die Anlagenhistorie ist ein Instrument, welches für die einzelnen Instandhaltungsobjekte die Informationen über Schadensbild, Schadensursache, die Beseitigungsmaßnahme und der geplante und tatsächliche Stundenaufwand für die Beseitigung beinhaltet. Die Aufzeichnungsdauer ist mit mehreren Jahren langfristig geplant.[392]

Andere Analysewerkzeuge sind der Soll-Ist-Vergleich, bei dem Vorgabewerte mit Ist-Werten verglichen werden. Das Ziel ist es Abweichungen zu identifizieren und Ursachen zu finden oder das Bilden und Auswerten von Kennzahlen.[393]

Alle Verfahren haben den Nachteil, dass sie bei einer größer werdenden Anzahl von Datenquellen und auszuwertenden Attributen nicht mehr geeignet sind. Daher müssen Methoden des Data-Mining eingesetzt werden und die Schwachstellenanalyse zu einer Big Data-gestützten Schwachstellenanalyse weiterentwickelt werden.[394]

5.2.2 Schwachstellenanalyse in der Instandhaltung

Die Instandhaltungsnorm DIN 31051:2012 unterscheidet die vier in Abbildung 26 dargestellten Grundmaßnahmen der Instandhaltung.

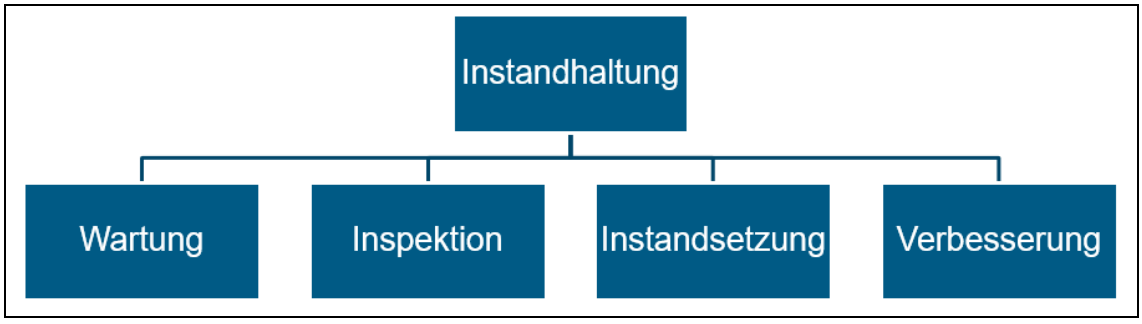

Abbildung 26: Grundmaßnahmen der Instandhaltung[395]

Die Verbesserung ist jene bei der die Schwachstellenanalyse als zentrale Methode zum Einsatz kommt. Die Norm sieht die Verbesserung als eine Kombination aus technischen und administrativen Maßnahmen, die eine Betrachtungseinheit verbessern, ohne deren

[391] Vgl. Bernerstätter, R.; Kühnast, R. (2017), S. 169 ff.
[392] Vgl. Biedermann, H. (1988), S. 316 ff.
[393] Vgl. Pawellek, G. (2016), S. 142-147.
[394] Vgl. Bernerstätter, R.; Kühnast, R. (2017), S. 170 f.; Biedermann, H. (2017), S. 35.
[395] Quelle: in Anlehnung an Deutsches Institut für Normung (2012), S. 4.

Funktion zu verändern. Die Verbesserungen können dabei die Zuverlässigkeit, die Instandhaltbarkeit oder die Sicherheit betreffen.[396]

Durch den verbessernden Charakter können die Maßnahmen, die in diesem Zusammenhang durchgeführt werden, mit der perfektiven Instandhaltung in Verbindung gebracht werden.[397] Ziel dieser Instandhaltungsstrategie ist es den Funktionsumfang zu verbessern bzw. zu erweitern. Dabei sollte diese Strategie nicht aus den anderen ausgewählt und statt diesen angewandt werden, sondern sie sollte laufend durchgeführt werden und andere Instandhaltungsstrategien ergänzen.[398] BIEDERMANN unterscheidet die anderen Strategien grob in die reaktive Instandhaltung als ausfallbasierende Strategie, die präventive Instandhaltung als ausfallvermeidende Strategie und die prädiktive Instandhaltung als ausfallprognostizierende Strategie. Wobei die prädiktive und die präventive Strategie noch in Substrategien unterteilt werden.[399]

Im Gegensatz zum klassischen Inhalt der Verbesserung laut der Norm, wird diese im Rahmen der Lean Smart Maintenance (LSM) Philosophie um die Dimension der Durchführung und der Organisation erweitert. Die Schwachstellenanalyse ist dabei ein wesentlicher Faktor die Ziele von LSM zu verwirklichen. Sie wird im Rahmen eines umfassenden Controllingsystems mit Unterstützung der Big Data Analytik dazu eingesetzt die Instandhaltungseffizienz zu steigern indem der Abnützungsvorrat erhöht, die Instandhaltbarkeit verbessert und die Durchführung optimiert werden.[400]

Abbildung 27 zeigt die gestärkte Rolle die die Schwachstellenanalyse im ursprünglichen Regelsystem[401] von BIEDERMANN spielt. Auf der operativen Ebene liefert die Schwachstellenanalyse Input für Aussagen über die Wirksamkeit von Instandhaltungsmaßnahmen. In Kombination mit einem Soll-Ist- Vergleich ist es möglich neben der generellen Wirksamkeit, die Effizienz der gesetzten Maßnahmen zu beurteilen. Die Ergebnisse fließen in die IH-Programmplanung ein. Darüber hinaus stellt sie Indikatoren zur Verfügung, die auf Over- oder Under-Maintenance hindeuten können. Damit kann eine Anpassung der Instandhaltungsstrategie einhergehen. Die Erfolgswirksamkeit der strategischen Anpassung kann mittelfristig mit der Schwachstellenanalyse überprüft werden. Durch eine Einbeziehung von Infrastruktur- und Umgebungsdaten im Sinne der vertikalen Integration eines MES ist es möglich den Wertschöpfungsbeitrag der Instandhaltung am Unternehmenserfolg darzustellen.[402]

Durch die verstärkte Digitalisierung, die auch in der Instandhaltung Einzug gefunden hat, ist es durch die horizontale und vertikale Integration der IT-Systeme möglich neue Datenquellen wie die BDE und MDE, in die Schwachstellenanalyse einzubeziehen.

Wird die CMMI-Systematik der Reifegradbewertung herangezogen, so haben die fünf Reifegrade für die Schwachstellenanalyse folgende Bedeutung:

1. Initial: Es gibt keine Schwachstellenanalyse (SSA)
2. Gemanagt: Die SSA wird unsystematisch durchgeführt

[396] Vgl. Deutsches Institut für Normung (2012), S. 6.
[397] Vgl. Warnecke, H.-J. (1992), S. 9.
[398] Vgl. Behrenbeck, K. R. (1994), S. 216 f.
[399] Vgl. Biedermann, H. (2018), S. 28 ff.
[400] Vgl. Biedermann, H. (2016), S. 20 ff.
[401] Siehe: Biedermann, H. (2008), S. 99.
[402] Vgl. Biedermann, H. et al. (2017), S. 14.; Biedermann, H. (2017), S. 32 f.

3. Definiert: Der SSA liegt ein standardisierter Prozess zugrunde
4. Quantitativ gemanagt: Es können organisatorische Schwachstellen gefunden werden
5. Optimierend: Es werden auf Datenanalyse basierte Ursache-Wirkungsketten erstellt, die aus horizontal und vertikale integrierten Informationssystemen die Daten beziehen

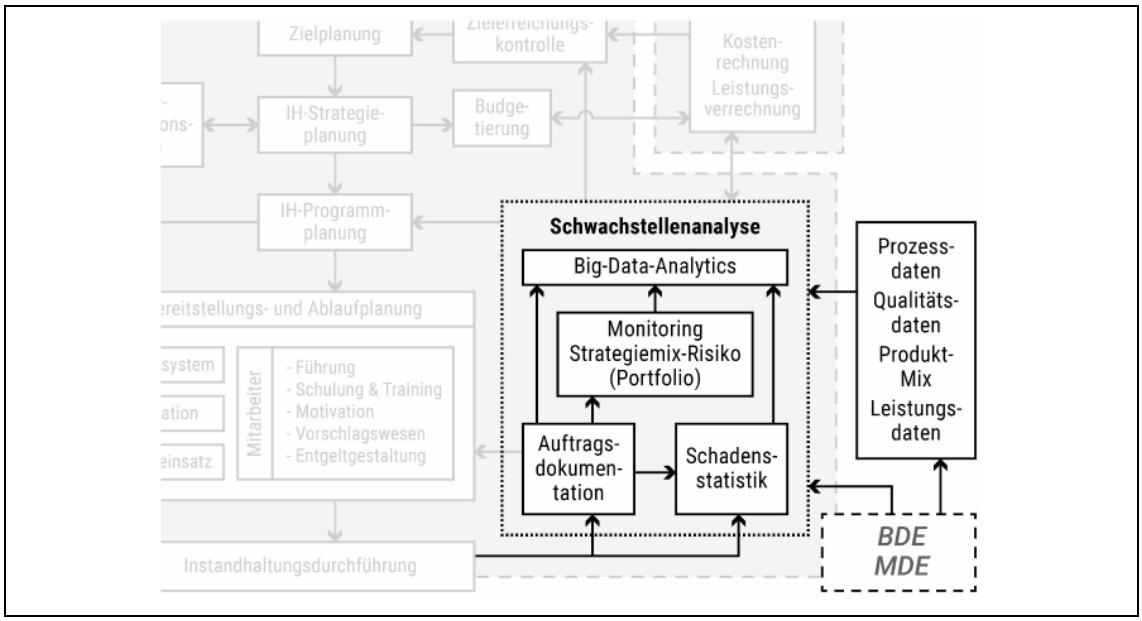

Abbildung 27: Schwachstellenanalyse im Instandhaltungscontrolling[403]

Schwachstellenanalysen ab dem Reifegrad vier entsprechen den Anforderungen die LSM stellt. Diese Art der Schwachstellenanalyse bedarf nicht nur einer Erweiterung des Begriffs der Schwachstelle, sondern auch neuer Methoden. Sie helfen die unterschiedlich anfallenden Daten zu nutzen und Schwachstellen zu finden, die aufgrund der steigenden Komplexität nicht mehr mit einfachen Methoden und den klassischen Daten zu finden sind.

5.2.3 Big Data gestützte Schwachstellenanalyse

In diesem Verfahren werden Daten aus unterschiedlichen Quellen mit einer Vielzahl an Attributen analysiert. Dabei soll auch die Ursachenfindung unterstützt werden indem diese Analysen es erlauben Ursache-Wirkungsketten aufzubauen.

Darin liegt der entscheidende Vorteil dieses Ansatzes, da Zustandsdaten ob der schieren Menge nicht bei klassischen deskriptiven Analysen mitbetrachtet werden.[404]

Störungsanalysen oder Ereignisablaufanalysen versuchen durch Diskussion und Erfahrungswerte Wirkungsketten aufzubauen. Ausgehend von der Störung wird basierend auf der Grundannahme, dass Fehler nicht ohne externe Einflüsse entstehen, nach möglichen Ursachen gesucht. Für diese wird wiederum nach deren Auslöser gesucht, bis man den Grundmangel gefunden hat.[405] Das Cause-Mapping geht ähnlich

[403] Quelle: Biedermann, H. (2017), S. 32.
[404] Vgl. Gartzen, T. et al. (2009), S. 48.
[405] Vgl. Strunz, M. (2012), S. 159 ff.

vor. Es werden jedoch bereits Datenaufzeichnungen entlang eines Zeitstrahls verwendet, um komplexe Netzwerke aus Wirkungen zu zeichnen. Dies erfolgt manuell mit unterschiedlichen Kreativitätstechniken. Den Ausgangspunkt der Analyse bildet die Verletzung eines Produktions-, Instandhaltungs- oder Sicherheitsziels.[406]

Beide Analysen setzten auf einen starken Input von Experten in der Erstellung der Wirkungsketten. Es besteht daher die Möglichkeit, dass Ursachen durch subjektive Einflüsse und vorgefasste Meinungen übersehen werden. Ziel sollte es sein nicht nur die bereits bekannten Probleme und Ursachen zu betrachten, sondern aufgrund der Datenfülle aus unterschiedlichen Quellen unbekannte Zusammenhänge zu finden. Ähnlich dem Vorgehen der Störungsanalyse und des Cause-Mappings wird mit Unterstützung der Assoziationsanalyse ein Vorgehen angewandt, welches die Auswertung automatisiert. Dieses Analyseverfahren ermöglich erst die Einbeziehung großer und hochdimensionaler Datenmengen.[407]

Mit den Interessantheitsmaßen Support, Konfidenz und Lift, werden die Regeln der Assoziationsanalyse konkretisiert. Abbildung 28 zeigt ein Big Data gestützte Ishikawa-Diagramm. Ausgangspunkt der Analyse ist jenes Problem, welches den Bedarf einer Schwachstellenanalyse ausgelöst hat. In der Abbildung wird diese als Endwirkung bezeichnet. Dabei kann es sich um einen häufig auftretenden Störcode oder ein unerklärliches Problem handeln. Deren Ursache und Auftreten ist mit herkömmlichen Methoden wie Kennzahlenprofile oder ABC-Analyse nicht identifizierbar. Sie werden als erste Konklusion {K} definiert. Die Ursachen der relevanten Regeln – bestimmt mit den Interessantheitsmaßen – werden in einem zweiten Schritt als Probleme definiert. Auf diese wirken die Nebenursachen. Die Schrittabfolge wird solange wiederholt, bis die Regeln nicht länger als relevant eingestuft werden können. Zum leichteren Verständnis können diese Zusammenhänge in einem Ishikawa-Diagramm visualisiert werden.[408]

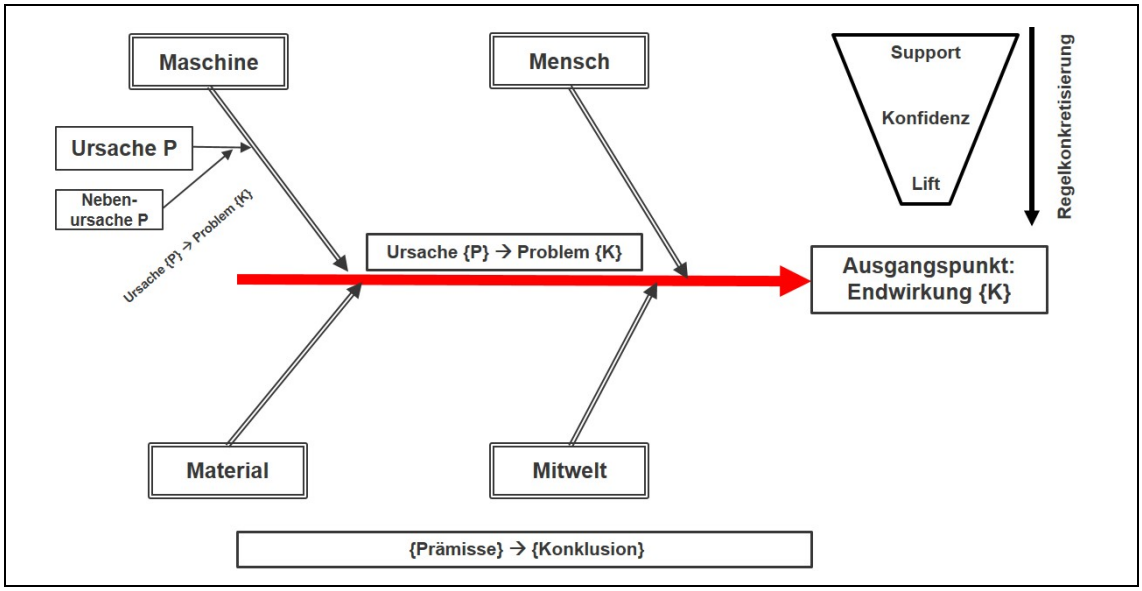

Abbildung 28: Big Data gestütztes Ishikawa Diagramm[409]

[406] Vgl. Horn, G. (2012), S. 12 ff.
[407] Vgl. Bernerstätter, R.; Kühnast, R. (2017), S. 171.
[408] Vgl. Bernerstätter, R.; Kühnast, R. (2017), S. 176 f.
[409] Quelle: in Anlehnung an Bernerstätter, R.; Kühnast, R. (2017), S. 177.

Mit dem Ishikawa Diagramm als Ergebnis, kann in eine Diskussion mit Domänenexperten in den betreffenden Prozessen getreten werden, um die aufgestellten Hypothesen zu den Zusammenhängen zwischen Ursache und Wirkung zu diskutieren. Das Verfahren und die Visualisierungsart wurden im Rahmen des DSR dieser Arbeit als Metaartefakt entwickelt.

5.3 Zusammenfassung und Relevanz für die Arbeit

Die systematische Analyse von unerwünschten Ereignissen in komplexen Systemen gewinnt durch die zunehmende Digitalisierung und die Integration flexibler und automatisierter Abläufe immer mehr an Bedeutung. Es greift daher zu kurz von einer reinen Schwachstellenanalyse zu sprechen, die sich nur auf Schäden von physischen Anlagenkomponenten konzentriert. Des Weiteren reicht es nicht sich einzig auf die Identifikation von Schwachstellen zu beschränken, sondern die Analyse muss bei der Suche nach den Ursachen unterstützen. Dabei kann sie auf die großen Datenmengen zurückgreifen, die durch die Integration entstehende Komplexität erzeugt werden. Es zeigt die Notwendigkeit Schwachstellen allgemeiner, über die Instandhaltung hinaus, zu definieren. Die Instandhaltung selbst stellt durch die horizontale und vertikale Integration nur einen Teil der Komplexität dar. Ursachen und Wirkungen beschränken sich nicht auf funktionale Grenzen, sondern sind selbst als Teil eines Fehlerprozesses zu betrachten. Die Definition der Schwachstelle sollte so generisch erfolgen, dass sie auf alle Bereiche anwendbar ist und von der Ursache abgegrenzt wird.

„Eine Schwachstelle ist eine an einer physischen oder nicht physischen (abstrakten) Messstelle feststellbare Wirkung die zu einer negativen Abweichung vom Sollzustand führt, welche durch zeitlich vorgelagerte Einflüsse (Ursachen) ausgelöst wurde. Die Ursachen sind dabei nicht zwingend örtlich – physisch wie organisatorisch – an die Messstelle gebunden.“

Die Schwachstellenanalyse hat speziell – aber nicht darauf beschränkt – in einer voll integrierten Industrie 4.0-Welt die Aufgabe die Zusammenhänge, die zwischen Ursache und Wirkung bestehen offenzulegen. Die Schwachstellenbeseitigung muss basierend darauf feststellen, wie diese Wirkungskette am besten unterbrochen wird, um eine technisch und/oder organisatorisch und wirtschaftlich günstige Lösung für das Problem zu erlangen, das die Schwachstelle darstellt. Um die Anforderungen an die Schwachstellenanalyse in komplexen Systemen zu erfüllen ist es nötig Daten und Datenmanagementsysteme zu haben, die Analysen auf dem nötigen Niveau ermöglichen. Wie in Abschnitt 5.2.2 beschrieben, steigen die Anforderungen an die Schwachstellenanalyse durch die fortschreitende Digitalisierung und die horizontale wie vertikale Integration. Die herkömmlichen Methoden aus Abschnitt 5.2.1 können diese nicht vollumfänglich erfüllen. Die ABC-Analyse beschränkt sich auf eine einzelne Variable, die auf Maschinenebene ausgewertet wird. Sie dient hauptsächlich dazu jene Komponenten zu identifizieren, die genauer betrachtet werden sollen. Kennzahlengestützte Methoden wie Kennzahlenvergleiche und –profile stellen im Gegensatz zur ABC-Analyse einen Verlauf von Kennzahlen dar. Die verwendeten Kennzahlen sind aggregierte Daten und die Verfahren daher nicht in der Lage Zusammenhänge zwischen den Kennzahlen festzustellen. Ähnliches gilt für die

Anlagenhistorie oder statistische Auswertungen. Die Kombination von unterschiedlichen Datenquellen und nicht aggregierten Daten als Input für die Instrumente ist nicht möglich. Außerdem ist es unmöglich Wirkungsketten und Kausalitäten abzubilden. Die kausalitätsbezogenen Analysen sind eine Möglichkeit diesen Nachteil zu kompensieren. Sie basieren jedoch auf vorab angestellten Überlegungen. Die Wirkungsketten und Szenarien können entweder nicht eintreten oder sie sind unvollständig und bilden die im Anlagenbetrieb auftretenden Szenarien nicht ab.

Die erwähnten Analysen sind in den geringsten beiden Reifegraden weiterhin anwendbar. In den beiden höchsten, ggf. den drei höchsten, Reifegraden werden Methoden und Algorithmen der Datenanalyse benötigt. Diese wurden in Abschnitt 4.2 beschrieben und in Abbildung 22 zusammengefasst. Es bedarf einer Methode die Anforderungen, die ein Problem an eine Schwachstellenanalyse oder eine allgemeine Datenanalyse stellt, mit den Möglichkeiten der Analyse abzugleichen. Darüberhinausgehend muss festgestellt werden, ob eine Organisation in der Lage ist, eine Analyse auf dem geforderten Niveau durchzuführen. So kann eine geeignete Methode für eine Schwachstellenanalyse gewählt und die Reife einer Organisation festgestellt werden. Zu diesem Zweck wurde das Reifegradmodell entwickelt, um darzustellen auf welchem Niveau sich ein Betrieb für komplexe oder weniger komplexe Analysen befindet. Die Schwachstellenanalyse mit Big Data-Unterstützung wird der diagnostischen Analyse zugeordnet. Sie ist bereits so komplex, dass höhere Anforderungen an die Daten gestellt werden, als für einfache deskriptive Analysen.

Für diese Arbeit zählen die Arten und Herangehensweisen der Schwachstellenanalyse aus Abschnitt 5.2.1 zu den einfachen deskriptiven Analysen, da sie für niederdimensionale Daten anwendbar sind. Analog zu Abbildung 23 geben sie nur Auskunft darüber was passiert ist. Analysen mit Big Data sollen Auskunft geben warum etwas passiert ist und werden der Diagnostik zugeordnet. Nach einer Identifikation von Mustern auf Basis der Diagnostik, können diese für Prognosen verwendet werden. Mit diesem ist es möglich festzustellen, wann etwas passiert um Schwachstellen vorab vermeiden. Hier handelt es sich um Analysen auf prognostischem Niveau.

Das Modell muss in der Lage sein diese Situationen in einen Reifegrad einzuordnen. Einerseits sind viele Unternehmen noch nicht soweit, um komplexe Data-Mining Methoden anzuwenden. Sie benötigen eine Aussage wo sie sich befinden und wie sie sich auf ein höheres Niveau entwickeln können. Dieses nächst höhere Niveau ist die Big Data gestützte diagnostische Schwachstellenanalyse. Andererseits sind für schnelle und einfache Analysen deskriptive Instrumente ausreichend.

Das Modell wird an Fallbeispielen von instandhaltungsrelevanten Schwachstellenanalysen ausgestaltet. Es wird dabei so vorgegangen, dass die oben gegebene Definition der Schwachstelle gilt und das Modell dadurch auf alle diagnostischen Analysen umlegbar ist. Als Beispiel für einfache Analysen gilt die ABC-Analyse und als Beispiel für Big Data gestützte Analysen die Assoziationsanalyse mit der Anwendungsart wie in Abschnitt 5.2.3 beschrieben. Auch wenn der theoretische Anspruch besteht, dass organisatorische Schwachstellen gefunden werden können, da diese mit der Schwachstellendefinition abgedeckt werden, werden diese explizit als Fallbeispiel ausgenommen. Sie können aufgrund der derzeit fehlenden Datenbasis in der Industrie durch Fallbeispiele nicht verifiziert werden.

6 Reifegradmodell zur Bewertung von Daten

Für die Bewertung einer Organisation hinsichtlich der Fähigkeit zur Umsetzung eines datenanalytischen Projektes wurde ein Reifegradmodell entwickelt, welches im folgenden Kapitel beschrieben wird. Die Ausgestaltung der Reifegradkategorien wurde im Rahmen der Durchführung von sechs Projekten vorgenommen. Drei der Projekte dienten der spezifischen Beschreibung anhand der Schwachstellenanalyse und gaben den Input für die Bewertung des Reifegrads durch Fragen und einfacher Analysen.

Die Entwicklung erfolgte entsprechend des Design Science Research (DSR) Vorgehens nach HEVNER[410]. Die prozessuale Basis bildet der CRISP-DM Zyklus. Es wurden sechs Reifegradkategorien definiert anhand derer die Reife der Organisation bewertet wird erfolgreich Datenanalysen durchzuführen.

Die sechs Reifegradkategorien sind:

1. Datenerfassung
2. Datenbereitstellung (Datenspeicherung und –übertragung)
3. Datenformate
4. Datendarstellung und -codierung
5. Datenumfang
6. Datenkonsistenz (zeitliche Konsistenz)

Die Reifegrade ergeben sich aus der Fähigkeit datenanalytische Fragestellungen beantworten zu können. Sie spiegeln die Bandbreite des Datenmanagements und der Datenstruktur von der rudimentären oder nicht vorhandenen Digitalisierung bis zur vollständigen horizontalen und vertikalen Integration wider. Es wurden vier generische Reifegradstufen definiert, die sich an den Analysearten aus Abbildung 23 orientieren.

BAUMÖL UND MESCHKE definierten für die Bewertung des Datenqualitätsmanagements vier Reifegrade. Die Benennung wird an CMMI angelehnt. Der Inhalt entspricht jedoch nicht den ursprünglichen Inhalten der Prozessreife laut CMMI. Die Reifegradstufen zeigen die steigende Komplexität des Datenqualitätsmanagements an. Der Übergang von Stufe zwei auf Stufe drei ist ein Übergang von einer auf Einzelebenen beschränktes Management auf eine horizontale Integration über Unternehmensbereiche hinweg.[411]

HEIDEL ET AL. definieren unterschiedliche Aspekte der Digitalisierung und Industrie 4.0 mit vier Abstufungen. Zum einen wird die Kommunikationsfähigkeit von Gegenständen und deren Identifikation in einem Netzwerk in vier Stufen kategorisiert.[412] Zum anderen werden Informationen und Daten nach deren Industrie 4.0 Konformität in vier Stufen unterteilt.[413]

[410] Vgl. Hevner, A. R. et al. (2004)
[411] Vgl. Baumöl, U.; Meschke, M. (2009), S. 65.
[412] Vgl. Heidel, R. et al. (2017), S. 34 f.
[413] Vgl. Heidel, R. et al. (2017), S. 51.

CAMM ET AL. oder DORSCHEL ET AL. unterteilen die Komplexität von Analytics-Methoden in drei bzw. fünf Stufen.[414] Die Unterteilung von CAMM ET AL. ist zu wenig granular, jene von DORSCHEL ET AL. enthält mit der Stufe des Monitorings eine unwesentliche Art für den datenanalytischen Prozess. Die Fähigkeit des Monitorings Daten in Echtzeit aufzuzeichnen und darzustellen wird in dieser Arbeit in die Stufe der Predicitve Analytics aufgenommen. Betrachtet man die möglichen Fragestellungen einer Datenanalyse, so sind die vier wesentlichen:

- Was ist geschehen?
- Warum ist etwas geschehen?
- Was wird geschehen?
- Was soll geschehen?

Diese vier Fragen benötigen steigende Komplexität der Analyse, um beantwortet zu werden. Damit gehen eine bessere Datenqualität und ein besseres Datenmanagement einher. Somit sind folglich die vier Reifegradstufen des Modells:

1. *Reifegrad* - Deskriptive Prozessreife: Visuelle Analysen und Berichterstellung
2. *Reifegrad* - Diagnostische Prozessreife: Korrelationen und Ursache-Wirkungen
3. *Reifegrad* - Prädiktive Prozessreife: Prognosemodelle mit hoher Güte
4. *Reifegrad* - Präskriptive Prozessreife: Autonome Entscheidungsfindung

In den folgenden Abschnitten werden die generelle Struktur des Reifegradmodells sowie die inhaltliche Ausgestaltung der Reifegradkategorien beschrieben. Dabei wird zuerst die theoretische Fundierung des Modells erörtert, um abschließend auf die Bewertung in den Kategorien mit den Reifegradstufen einzugehen.

6.1 Struktur des Reifegradmodells

Das Reifegradmodell wurde mit dem DSR-Vorgehen entwickelt. Das grundlegende Design des Modells wurde im Laufe mehrerer Projekte verfeinert. Abbildung 32 zeigt die finale Version der Struktur und des Aufbaus, der Ablauf und die Einbettung in den Bewertungsprozess ist in Abbildung 34 zu sehen. Zur methodischen Bestimmung der Struktur des Reifegradmodells wurde das Quality Function Deployment verwendet. Damit soll der Zusammenhang zwischen den Prozessphasen der Datenanalyse und den Reifegradkategorien hergestellt werden.

Quality Function Deployment (QFD) und House of Quality (HoQ)

QFD wurde in den 1970er Jahren in Japan von Prof. Yoji Akao entwickelt. Starke Verbreitung fand die Methode in der Automobilindustrie. Es handelt sich bei der Methode, um ein strukturiertes Vorgehen, die Qualitätsplanung auf den Kundenanforderungen an Produkte und Prozesse aufzubauen. Ziele sind u. a. die Erfüllung der Kundenanforderungen bzw. die Verbesserung des Erfüllungsgrades, die

[414] Vgl. Dorschel, J. et al. (2015), S. 56 f.; Camm, J. D. et al. (2018), S. 11.

Motivation zum Mitgestalten und Mitdenken zu fördern, die Schaffung abgestimmter Ziele und die Verbesserung der Wettbewerbsposition.[415]

Das House of Quality (siehe Abbildung 29) ist das Werkzeug, welches das QFD unterstützt. Es hilft bei der Beantwortung der Fragen:[416]

- Was erwartet der Kunde? – Was?
- Wie werden die Erwartungen erfüllt? – Wie?
- Wie stark müssen die Ausprägungen erfüllt werden? – Wieviel?
- Wie gut erfüllt der Wettbewerb die Erwartungen des Kunden? – Warum?

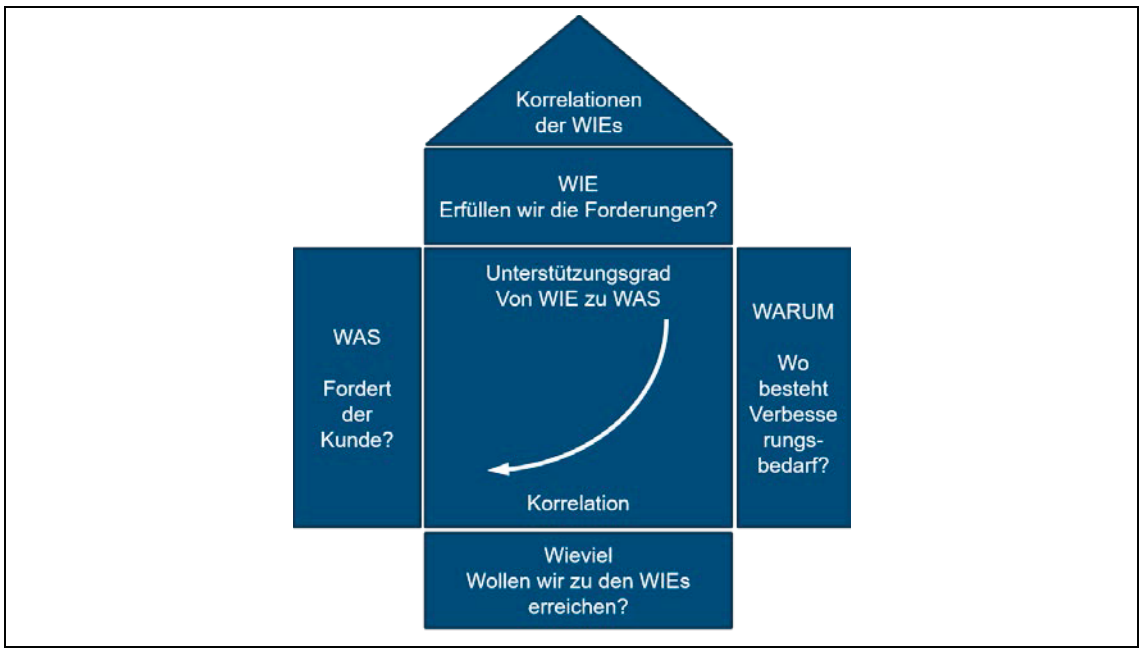

Abbildung 29: House of Quality[417]

Umgelegt auf die Anwendung im Reifegradmodell und den CRISP-DM sind die Kundenanforderungen durch das Unternehmensziel und die datenanalytische Problemstellung gegeben. Bei „Wie“ handelt es sich um die technischen Spezifikationen, die durch die Reifegradkategorien abgebildet werden. Die Datenqualitätsdimensionen spiegeln das „Wieviel“ wider. In der Korrelationsmatrix des Unterstützungsgrades wird gezeigt, wie der Zusammenhang zwischen den Reifegradkategorien und den Prozessschritten des CRISP-DM ist. Im Dach des Hauses werden die Korrelationen zwischen den Reifegradkategorien festgestellt. Die Leistungsvergleiche erfolgen im „Warum“ und im „Wieviel“. Hierfür wird die Bewertung mit dem Reifegradmodell herangezogen. [418] Es wird dabei der SOLL und der IST Reifegrad verglichen, bzw. eine Messung mit den Datenqualitätsdimensionen vorgenommen. Die Ergebnisse sind ein IST-Reifegrad des Unternehmens in Bezug auf den datenanalytischen Prozess und Handlungsempfehlungen zur Erreichung eines SOLL Reifegrades.

[415] Vgl. Kiem, R. (2016), S. 175.
[416] Vgl. Linß, G. (2018), S. 284.
[417] Quelle: in Anlehnung an Linß, G. (2018), S. 285.
[418] Vgl. Linß, G. (2018), S. 284 f.

House of Data Quality

Analog zum House of Quality wurde im Zuge dieser Arbeit ein House of Data Quality (HoDQ) erarbeitet, welches Abbildung 30 zeigt. Es zeigt welche Reifegradkategorien für welche Prozessschritte der Datenanalyse relevant sind. Darüber hinaus wird der Konnex zwischen den Datenqualitätskategorien und der Datenanalyse hergestellt.

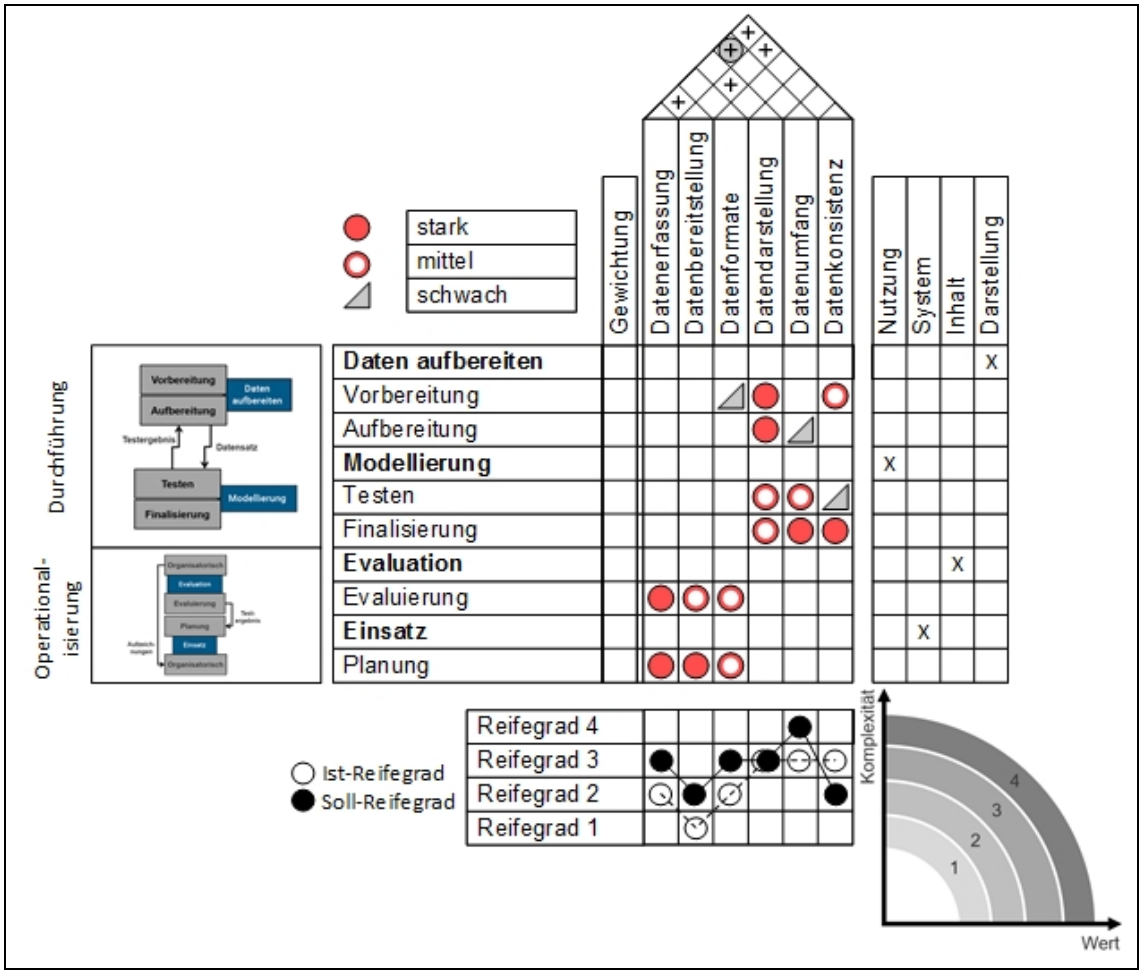

Abbildung 30: House of Data Quality[419]

Auf der linken Seite sind die CRISP-DM Phasen mit den aggregierten Prozessschritten abgebildet. Die Zusammenfassung der Schritte entspricht jener aus Tabelle 11. Die Abbildungen in den Rechtecken Durchführung und Operationalisierung sind Ausschnitte aus Abbildung 35 und Abbildung 36. Deren Inhalt wird in Abschnitt 6.2.2 genauer beschrieben. Die Matrix aus den Spalten der Reifegradkategorien und den Zeilen der Prozessschritte zeigt die Wichtigkeit der Kategorien in Bezug auf den Datenanalyseprozess. Die Bewertungen – stark, mittel, schwach – ergaben sich aus der wiederholten Anwendung in den sechs Fallbeispielen. Diese werden in Kapitel 7 detailliert beschrieben.

[419] Quelle: Eigene Darstellung

Die Datenerfassung ist entscheidend für die Evaluation und die Planung des Einsatzes des erstellten Modells. In diesem Schritt wird bestimmt, ob eine spätere Implementierung des Modells möglich ist.

Die Datenbereitstellung ist wichtig für die Operationalisierungsphase des CRISP-DM. Die Art der Datenarchitektur und –modellierung spielt hier eine Rolle. Die Bereitstellung von Daten ist wichtig für eine reibungslose Integration des Modells in den laufenden Geschäftsbetrieb.

Die Kategorie der Datenformate ist von mittlerer Wichtigkeit für die Operationalisierungsphase. Eine Kombination mehrerer Datenformate ist im Echtzeitbetrieb aufwendig aufzubereiten und daher für den Einsatz bei zeitkritischen Aufgabenstellungen unmöglich. Ungünstige Datenformate erschweren die Vorbereitung zur Datenaufbereitung, weil Formate erst transformiert werden müssen.

Die Datencodierung behandelt das Skalenformat in dem Daten abgebildet sind bzw. ob sie strukturiert oder unstrukturiert vorliegen. Die gesamte Durchführungsphase wird durch diese Kategorie beeinflusst. Sind die Daten unstrukturiert, müssen Informationen aufwendig extrahiert werden, um sie für die spätere Modellierungsphase zugänglich zu machen. Für die Modellierung ist diese Kategorie wichtig, da unterschiedliche Skalenformate unterschiedliche Analysealgorithmen erlauben.

Der Datenumfang ist besonders relevant für die Modellierung. Die Datenmenge erlaubt es maschinellen Lernalgorithmen allgemein gültige Muster zu erkennen. Zur Datenaufbereitung besteht ein schwacher Einfluss. Ein geringer Datenumfang kann die Datenaufbereitung vor die Herausforderung stellen Daten oder Merkmale zusätzlich zu erzeugen.

Die zeitliche Konsistenz schränkt den Einsatz von Methoden zur Darstellung von Sequenzen und Abläufen ein. Eine schlechte Konsistenz der Daten, speziell auf zeitlicher Ebene, erfordert unter Umständen aufwendige Aufbereitungsschritte, um die Konsistenz der Daten herzustellen.

Die Beeinflussung der Reifegradkategorien untereinander wird im Dach des HoDG dargestellt. Die Datenerfassung beeinflusst die Kategorie der Datenbereitstellung dadurch, dass die Erfassungsart die Art der Datenquelle beeinflussen kann. Eine indirekte Rolle in der Datenaufbereitung spielt die Datenerfassung über den Einfluss auf die Kategorie des Datenumfangs und der Datenkonsistenz. Automatische Erfassungssysteme sind in der Lage größere Datenmengen zu erfassen als manuelle Auszeichnungen. Daher beeinflusst die Datenerfassung die Kategorie des Datenumfangs. Der Zusammenhang zwischen Datenbereitstellung und Datenkonsistenz besteht, da Daten aufgrund der bestehenden Schlüsselbeziehung in einer gemeinsamen Quelle eher abgestimmt sind. Die Datendarstellung wird durch die Datenbereitstellung beeinflusst, da integrierte Systeme eher eine einheitliche Darstellung und eindeutige Auslegbarkeit der Daten garantieren.

Die rechte Seite des HoDG in Abbildung 30 zeigt die Zuordnung der Datenqualitätskategorien zu den Prozessschritten des CRISP-DM. Über die Korrelationsmatrix aus Reifegradkategorien und Prozessschritten wird die Beziehung zwischen den Datenqualitätskategorien und den Reifegradkategorien hergestellt. Die Datenqualitätsbetrachtung bietet die Möglichkeit der objektiven Bewertung über

Datenqualitätsmetriken. Die genaue Beschreibung der Datenqualitätskategorien im Reifegradmodell erfolgt in Abschnitt 6.2.3.

Abbildung 31 zeigt die Aufgabe der Vorbereitungsphase im Reifegradmodell. Die Darstellung im linken Bereich ist ein Ausschnitt aus Abbildung 33 wobei das Vorgehen in der Bewertung in Abbildung 34 gezeigt wird.

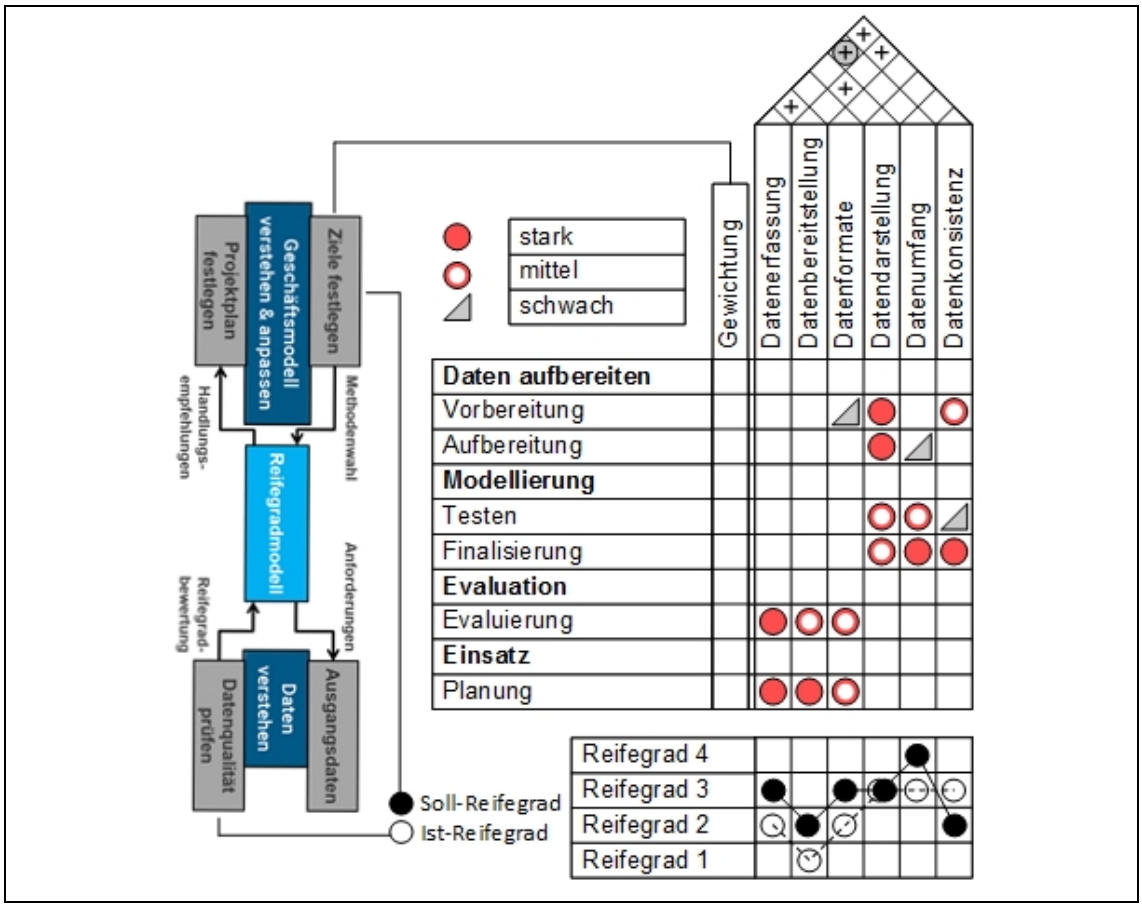

Abbildung 31: Vorbereitungsphase im HoDG[420]

Die Zielfestlegung bestimmt die Gewichtung der Prozessschritte und der Reifegradkategorien sowie in weiterer Folge den Soll-Reifegrad. Die Bewertung des Reifegrades in den Kategorien ergibt die Ist-Reifegradkurve. Diese wird mit einer Soll-Reifegradkurve verglichen. Die negative Abweichung (Soll<Ist) führt zu Handlungsempfehlungen, um einen höheren Reifegrad und folglich das datenanalytische Ziel zu erreichen.

Das Ergebnis der Feststellung der Zusammenhänge zwischen Reifegradkategorien, Prozessschritten, Datenqualitätskategorien und der Bewertung ist in Abbildung 32 zu sehen. Der Aufbau ist angelehnt an die Theorie aus Abschnitt 2.3.2, Abbildung 4 und Abbildung 5. Die Struktur des Reifegradmodells besteht aus drei übergeordneten Ebenen, die durch sieben untergeordnete Ebenen strukturiert werden. Übergeordnet sind die theoretische Ebene (Abschnitt 6.2), die empirische Ebene (Abschnitt 6.3) und die Bewertungsebene (Abschnitt 6.4). Der theoretischen Ebene untergeordnet sind die

[420] Quelle: Eigene Darstellung

Prozessfestlegung (Abschnitt 6.2.1), die Prozessphasen (Abschnitt 6.2.2) und die Datenqualität (Abschnitt 6.2.3). Die Reifegradkategorien (Abschnitt 6.3.1) und die Erhebungsebene (6.3.2) sind der empirischen Ebene untergeordnet. Die Reifegradstufen (Abschnitt 6.4.1) und die finale Bewertung (Abschnitt 6.4.2) gehören zur Bewertungsebene.

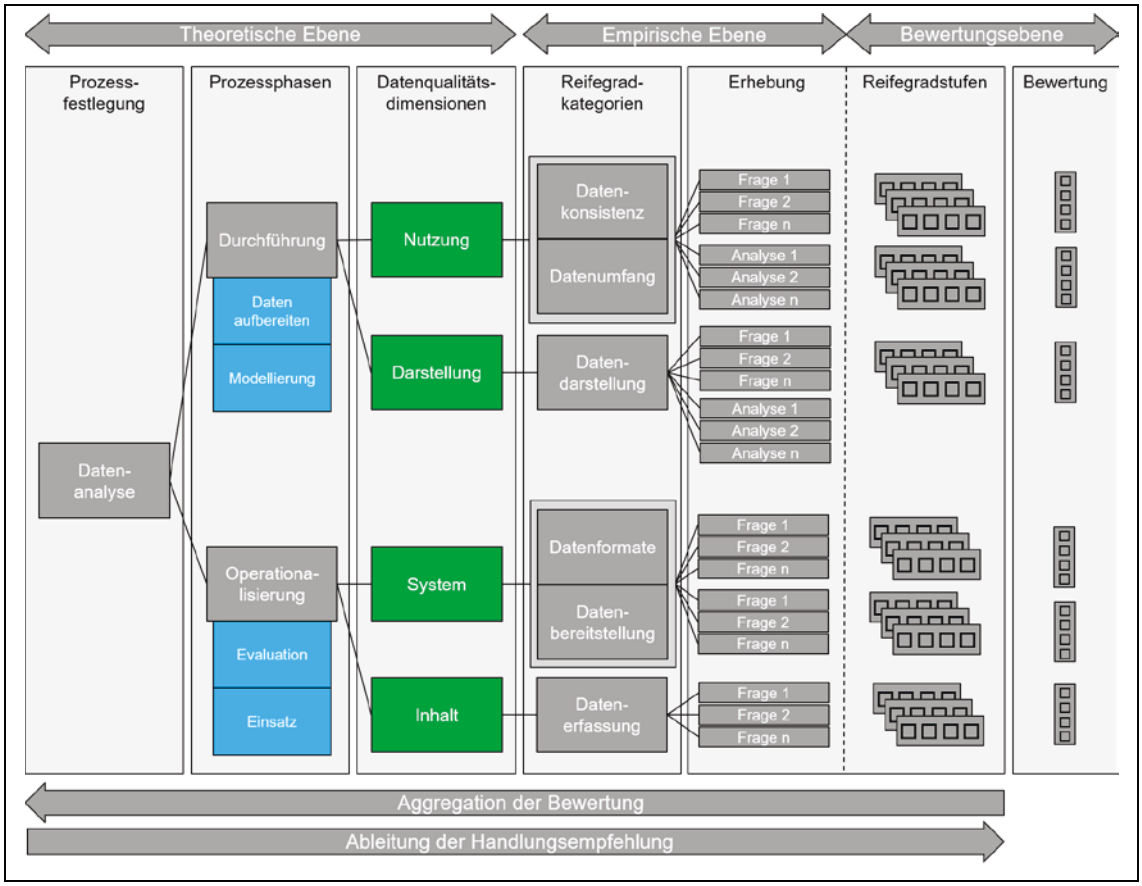

Abbildung 32: Aufbau und Struktur des Reifegradmodells[421]

Die Handlungsempfehlungen für die nötigen Verbesserungsmaßnahmen ergeben sich aus den Ebenen der Reifegradstufen und der Erhebung. Ein Vergleich von IST zu SOLL erlaubt es Maßnahmen zu definieren, um die Lücke zu schließen und den Reifegrad zu erhöhen.

6.2 Theoretische Ebene

Die theoretische Ebene bildet das Fundament zur Spezifizierung des Prozesses, der zur Bewertung steht. Die Ebene untergliedert sich in die drei Subebenen: Prozessfestlegung, Prozessphasen und Datenqualitätsdimensionen. Der Inhalt der theoretischen Ebene beschreibt, wie und wieso das Reifegradmodell den CRISP-DM und die Datenqualitätsdimensionen für die Bewertung in der empirischen Ebene verwendet.

[421] Quelle: Eigene Darstellung

Der zugrundeliegende Prozess der Datenanalyse ist der CRISP-DM. Dessen sechs Phasen – Geschäftsmodell verstehen, Daten verstehen, Daten aufbereiten, Modellierung, Evaluation und Einsatz – wurden in drei Metaphasen – Vorbereitung, Durchführung und Operationalisierung – zusammengefasst. Abbildung 32 zeigt die Beziehungen zwischen den Prozessphasen und Abbildung 33 und Abbildung 34 zeigen die Aufgabe der Vorbereitungsphase im Bewertungsprozess.

6.2.1 Prozessfestlegung

Die Ebene der Prozessfestlegung definiert das datenanalytische Problem. Sie bildet wie in Abbildung 32 gezeigt wird mit dem zu bewerteten Prozess der Datenanalyse die Basis der weiteren Bewertung. Hier werden die Geschäftsziele festgelegt aufgrund derer die Data-Mining Ziele festgelegt werden. Der Austausch zwischen den beiden Phasen der Vorbereitung, läuft über das Reifegradmodell. Abbildung 33 zeigt die Rolle des Reifegradmodells im CRISP-DM während Abbildung 34 das Vorgehen der Reifegradbewertung in der Ebene Prozessfestlegung zeigt.

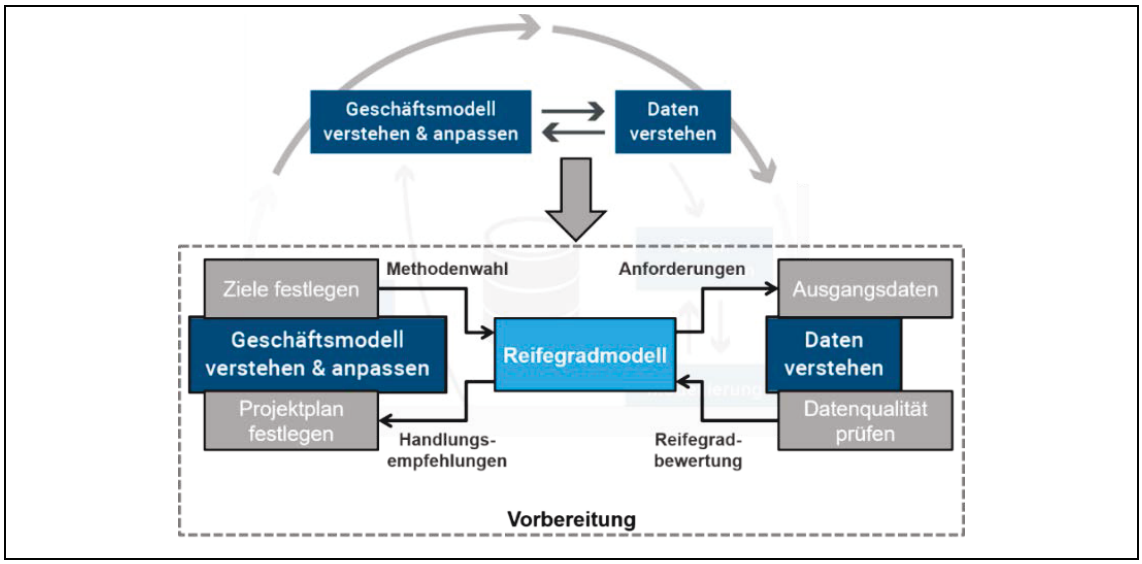

Abbildung 33: Reifegradmodell im CRISP-DM Ablauf[422]

Nach der Festlegung der Ziele erfolgt die Bewertung des Prozesses „Datenanalyse“. Dieser besteht aus den beiden Metaprozessphasen „Durchführung“ und „Operationalisierung“. Die Bewertung erfolgt auf der empirischen Ebene des Reifegradmodells durch die sechs Reifegradkategorien.

Die Anforderungen an die Durchführung und die Operationalisierung erlauben es einen SOLL-Reifegrad festzulegen. Der sich ergebende IST-Reifegrad wird damit verglichen. Ist der SOLL-Reifegrad höher, so werden Handlungsempfehlungen zur Reifegradverbesserung gegeben. Danach wird der Projektplan festgelegt und die Datenanalyse wird gemäß der CRISP-DM Phasen abgewickelt.

[422] Quelle: Eigene Darstellung

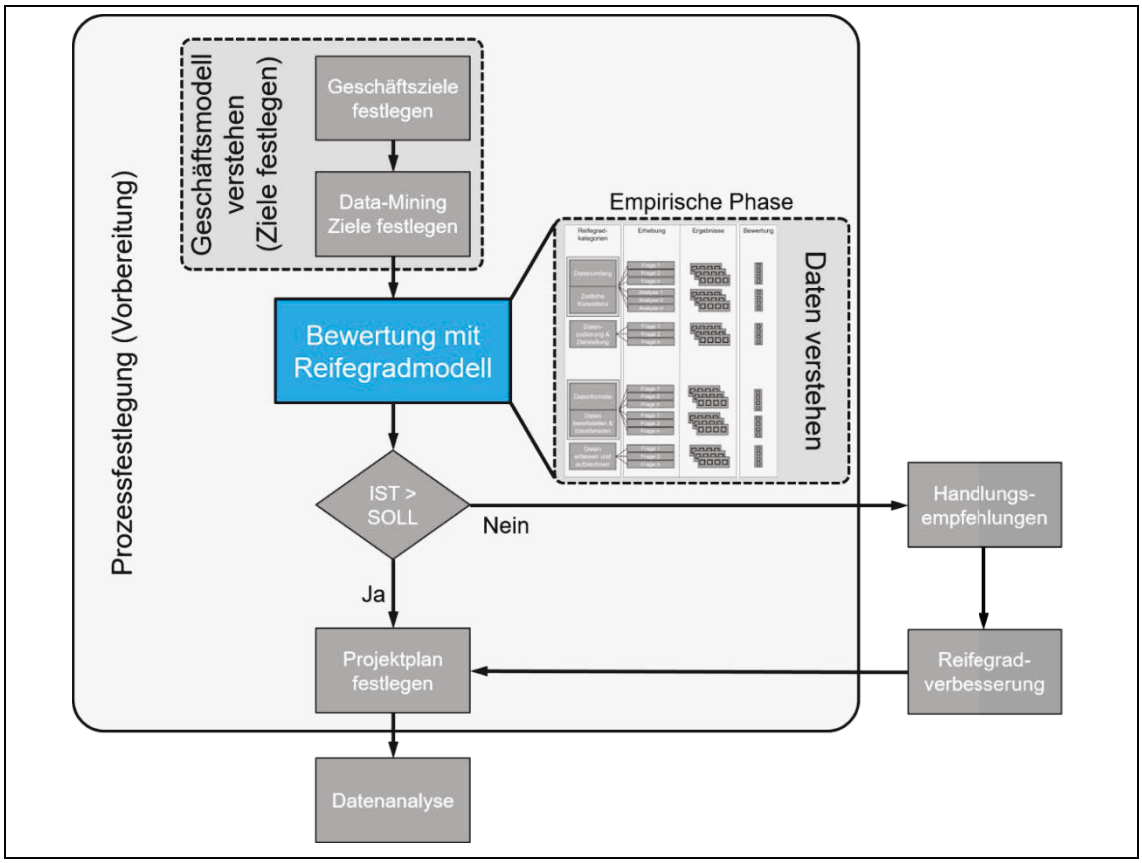

Abbildung 34: Vorgehensmodell des Einsatzes des Reifegradmodells[423]

Ziele festlegen

In der Phase „Geschäftsmodell verstehen“, werden die Geschäftsziele des Projektes festgelegt und diese in Data-Mining Ziele überführt. Somit ist es möglich, wie in Abbildung 22 angedeutet, die Aufgabenart der Analyse und eine Vorauswahl des Data-Mining Verfahrens zu treffen.[424]

Es lassen sich drei Fälle für den Einsatz definieren, die mit steigender Komplexität und Autonomiegrad die Gewichtung der Operationalisierungsphase erhöht. Die Reife in den zugeordneten Bewertungskategorien muss höher sein, da manuelle Nacharbeiten nicht mehr möglich sind.:

1. Einmalige Umsetzung durch einen Spezialisten.
    - Zum Beispiel eine Machbarkeitsstudie oder eine einmalige Beantwortung einer Fragestellung.
2. Mehrfache Umsetzung des Vorgehens entweder durch einen Spezialisten oder eine fachkundige unterwiesene Person.
    - Wiederholende Analysen in der Berichterstellung oder regelmäßige Diagnosen.
3. Automatische Implementierung der Durchführungs- und Operationalisierungsphase.
    - Die automatische Vorhersage von Zielfunktionen oder die autonome Ausführung von Aktionen.

[423] Quelle: Eigene Darstellung
[424] Vgl. Dippold, R. et al. (2005), S. 243 f.

Im ersten Fall geht es um eine einmalige Vergangenheitsbetrachtung. Es ist nicht das Ziel das entwickelte Modell regelmäßig einzusetzen.

Der zweite Fall hat zum Ziel die Durchführung mehrmals zu wiederholen. Entweder der gesamte Prozess, von der Datenaufbereitung über die Modellierung bis hin zur Operationalisierung, oder nur die Anwendung des einmal durch einen Spezialisten erstellten Modells auf neu angefallene Daten.

Der dritte Fall hat die höchsten Anforderungen an die Operationalisierung. Das Modell und das Vorgehen werden in den laufenden Geschäftsbetrieb implementiert. Die Planung der Datenaufbereitung und die Modellerstellung erfolgen auch hier durch einen Spezialisten. Die Anwendung ist jedoch nicht einmalig oder regelmäßig sporadisch, sondern voll automatisiert und in Echtzeit.

Die Data-Mining Ziele bzw. die Problemstellung ergibt sich aus der Geschäftszielen des Projektes. Wie bereits in den Abschnitten 4.2.2 und 4.2.3 erwähnt, haben die Methoden und Algorithmen unterschiedliche Voraussetzungen an die Daten. Diese sind durch die Data-Mining Ziele festgelegt.

Tabelle 12 gibt einen Überblick über die Data-Mining Methoden und die passenden Algorithmen zusammen mit möglichen Anwendungsbeispielen. Die Algorithmen sind im Zuge der Anwendungsbeispiele in den entsprechenden Quellen näher erörtert. Die Data-Mining Problemstellung für die Verfahren wurde in Abschnitt 4.2.1 kurz erörtert.

Eine Zuordnung von Geschäftszielen und Analyseverfahren wurden von KOPCSO UND PACHAMANOVA vorgenommen. Dabei wurden die Geschäftsziele entsprechend der Analysekomplexität gereiht.[425]

Tabelle 12: Übersicht Data-Mining Ziel und Projektziel[426]

Data-Mining Ziel	Verfahren	Projektziel
Klassifikation	Entscheidungsbaum	• Klassifikation von Anlagefehlern[427]
	Support Vector Machine	• Identifikation von mehreren Anlage-Fehlerklassen[428]
Prognostizieren	Lineare Regression	• Prädiktive Instandhaltung[429] • Restlebensdauerabschätzung[430]
	Zeitreihenanalyse	• Restlebensdauerabschätzung[431]

[425] Vgl. Kopcso, D.; Pachamanova, D. (2018), S. 43 f.
[426] Quelle: Eigene Darstellung
[427] Vgl. Nemeth, T. et al. (2015), S. 572.
[428] Vgl. Yan, J. (2015), S. 82 ff.
[429] Vgl. Bernerstätter, R. (2018a), S. 136 f.
[430] Vgl. Bernerstätter, R. et al. (2016), S. 27 f.
[431] Vgl. Yan, J. (2015), S. 162 f.

Fortsetzung Tabelle 12: Übersicht Data-Mining Ziel und Projektziel[432]

Muster erkennen	Clustern k-means	• Anlagenrisiken gruppieren[433]
	Clustern hierarchisch agglomerativ	• Verbesserung der Instandhaltungsplanung[434]
Anomalieerkennung	Multivariat	• Anlagenüberwachung[435]
	Univariat	• Identifizieren defekter Sensoren[436]
Zusammenhänge erkennen	Assoziationsanalyse	• Ersatzteilmanagement[437] • Störcodeanalyse[438] • Schwachstellenanalyse[439] • Verbesserung der Instandhaltungsplanung[440] • Produktionsoptimierung[441]
	Sequenzanalyse	• Schwachstellenanalyse[442]
Variablen reduzieren	Hauptkomponenten-analyse	• Finden der wichtigsten Einflussfaktoren. Z. B. wichtigste Sensoren

6.2.2 Prozessphasen

Der Datenanalyseprozess wird auf Ebene der zwei Metaprozessphasen, der Durchführung und der Operationalisierung, bewertet (siehe Abbildung 32). Die Inhalte der Phasen basieren auf dem CRISP-DM Modell. Abbildung 35 zeigt die Durchführungsphase und Abbildung 36 die Operationalisierungsphase.

Durchführung

In der Durchführung werden die wesentlichen datenanalytischen Schritte umgesetzt. Sie umfasst die spezifischen Handlungen und Ziele die für die Datenanalyse abgearbeitet und erreicht werden müssen. Der Input für die Durchführung kommt von der Vorbereitungsphase. Dort wurden die Geschäftsziele und die Data-Mining Ziele festgelegt. Basierend darauf konnten die Art der Analyse und der Analysealgorithmus gewählt werden. Eine Auswahl ist Tabelle 12 zu entnehmen.

[432] Quelle: Eigene Darstellung
[433] Vgl. Bernerstätter, R. et al. (2016), S. 39 f.
[434] Vgl. Yan, J. (2015), S. 215 ff.
[435] Bernerstätter, R.; Hirschmugl, R. (2018)
[436] Vgl. Deeskow, P. et al. (2008), S. 3.
[437] Moharana, U. C.; Sarmah, S. P. (2015)
[438] Vgl. Kleindienst, B.; Bernerstätter, R. (2015), S. 172 ff.
[439] Vgl. Bernerstätter, R.; Kühnast, R. (2017), S. 171 ff.
[440] Mosaddar, D.; Shojaie, A. A. (2013)
[441] Vgl. Kamsu-Foguem, B. et al. (2013), S. 1036 ff.
[442] Vgl. Bernerstätter, R.; Kühnast, R. (2017), S. 173 ff.

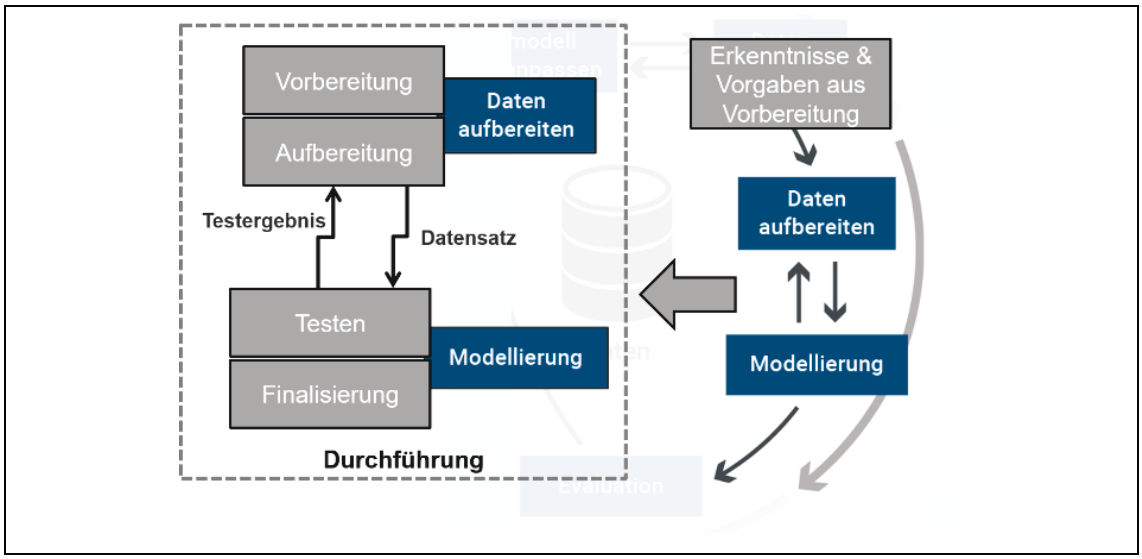

Abbildung 35: Metaphase Durchführung[443]

Die CRISP-DM Phasen in der Durchführung sind „Daten aufbereiten“ und „Modellierung“. Zwischen ihnen herrscht ein laufender Austausch, um die Ziele der datenanalytischen Problemstellung zu erreichen. Die Prozessschritte in der Datenaufbereitung können im Wesentlichen in jene mit Unterstützungsfunktion, die der Vorbereitung dienen und jener mit der zentralen Aufbereitungsfunktion unterteilt werden.

Daten aufbereiten:

Die Vorbereitung umfasst die Schritte der Datenauswahl und der Datenintegration. Die Datenintegration benötigt abhängig von der Datenhaltung und den vorherrschenden Formaten merklich manuellen Aufwand. Sie werden von der Wahl der Analysemethode und des Algorithmus wenig beeinflusst. Die Aufbereitung beinhaltet Schritte, die die Qualität und Struktur der Daten beeinflussen und für die Modellierung vorbereiten. Die Schritte sind die Datenbereinigung, die Erstellung neuer Daten und die Formatierung der Daten für das erforderliche Skalenniveau des Analysealgorithmus. Die Abfolge der drei Schritte muss nicht zwingend in dieser Reihenfolge sein.

Die Datenbereinigung verwendet die Erkenntnisse der Vorbereitungsphase. Die Handlungsempfehlungen ergeben sich aus den Voranalysen der Erhebungsebene der Reifegradkategorien.

Die Erstellung neuer Daten wird stark von der vorhandenen Datenbasis beeinflusst. Das Reifegradmodell beurteilt die Möglichkeit neue Daten zu erstellen. Die stärkste Beeinflussung ergibt sich aus dem Skalenniveau. Neue Daten werden durch die Kombination vorhandener Daten oder durch die Anwendung mathematischer Operationen erzeugt.[444] Die einsetzbaren mathematischen Operationen hängen direkt vom Skalenniveau der Daten ab. Ähnliches gilt für die Datentransformation. Diese hängt zum einen vom gewählten Algorithmus ab. Dieser gibt vor, welches Skalenniveau nötig ist und folglich welche Transformation im Schritt Datentransformation durchgeführt

[443] Quelle: Eigene Darstellung
[444] Vgl. Huang, J.; Esbensen, K. H. (2000); Cios, K. J. et al. (2007), S. 133 ff.; Schenkendorf, R.; Böhm, T. (2015), S. 17 ff.

werden muss. Zum anderen hängt die Möglichkeit der Transformation direkt mit dem Skalenniveau zusammen.

Modellierung:

Die Modellierung umfasst die Schritte der Testphase und der Finalisierung. Die Testphase umfasst jene Schritte, die für die Modellierung nötig sind, wie die endgültige Wahl des Modellierungsverfahrens und des Algorithmus, basierend auf den Daten der Datenaufbereitung und die Erstellung eines Testdesigns. Das Testdesign dient für die spätere Validierung des Ergebnisses und für die erste Abschätzung ob das Modellierungsverfahren und die Daten zu einem zufriedenstellenden Ergebnis führen können. Die Menge der Daten und das Skalenniveau beeinflussen die Wahl des Modells und das Testdesign stark.

Die Finalisierung der Modellierung ist die Erstellung des Modells und dessen Bewertung. Es handelt sich um die letzten Schritte der Durchführungsphase und bilden den Abschluss der Kernaufgabe eines Data-Mining Projektes. Das erzeugte Modell und die generierten Ergebnisse sind nach der Durchführung der Prozessschritte der bestmögliche Output. Die Bewertung, die mit dem Reifegradmodell durchgeführt wurde, sollte sicherstellen, dass die Phasen, die bis hier durchgelaufen wurden, nur bei einer ausreichend hohen Reife bearbeitet werden.

Operationalisierung

Die Operationalisierung beschäftigt sich mit der Umsetzung der Ergebnisse der Durchführungsphase. Dabei werden die Vorgaben und Erkenntnisse aus der Vorbereitung berücksichtigt. Daher ist es wichtig - wie in Abbildung 32 angedeutet - , dass diese Phase in der Reifegraderhebung berücksichtigt wird. Die Phase umfasst die CRISP-DM Schritte der Evaluation und des Einsatzes. Des Weiteren wird mit der Operationalisierung auch der Aufwand der Analyse angesprochen.

Evaluation:

Die Evaluation bewertet das Ergebnis der Durchführung, welches in Form des Modells, der Modellparameter und der Testergebnisse vorliegt. Die Evaluierung des Analyseprozesses ist von organisatorischer Natur und wird durch das Reifegradmodell nicht bewertet. Während das Augenmerk in der Durchführungsphase in der Bewertung des Modells in datenanalytischer Sicht liegt, wird in dieser Phase die Zielerfüllung aus der Projektsicht beurteilt. Daher sind die vorab definierten Ziele und das Domänenwissen der Experten ein wichtiger Input in dieser Phase.

Die Evaluierung und die Planung der nächsten Aktionen erfolgt im Evaluierungsschritt. Bis zu einem gewissen Grad haben alle Kategorien des Reifegradmodells Einfluss. Die Evaluierung der Ergebnisse hängt von der Datenmenge und der Glaubwürdigkeit der Quellen ab.

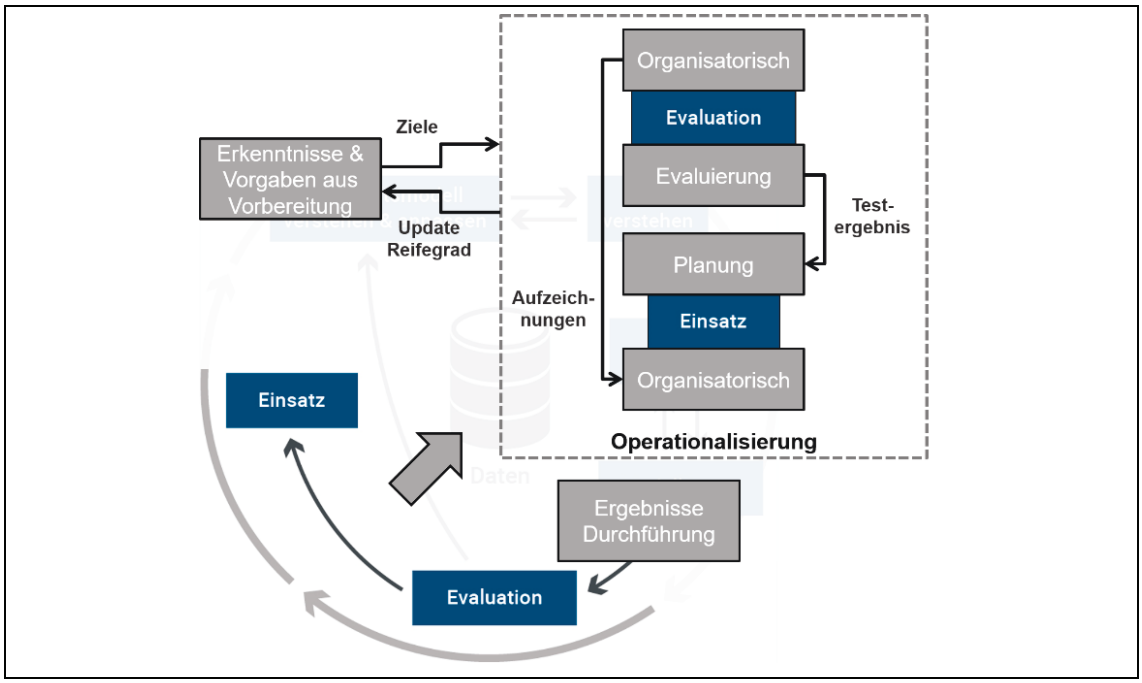

Abbildung 36: Analysephase Operationalisierung[445]

Einsatz:

Falls die Entscheidung zum Einsatz des Ergebnisses der Durchführung getroffen wird, hängt die Art der Implementierung vom Reifegrad in den relevanten Kategorien ab. Der Prozessschritt des CRISP-DM, der Einsatz, wurde in die Teilschritte Planung und jene Aktionen, die von organisatorischer Natur sind, unterteilt. Die Letzteren werden durch das Reifegradmodell nicht bewertet.

Die Schritte der Planung, die eigentliche Planung und die Planung der Überwachung und Wartung des Modells, hängen von ihren Möglichkeiten stark am Reifegrad der beiden ersten Kategorien des Modells ab. Beide CRISP-DM Schritte Evaluation und Einsatz erhalten die Ziele und Vorgaben als Input aus der Vorbereitung. Eine etwaige Aktualisierung des Reifegrades wird zur Vorbereitung zurückgespielt. Entweder durch eine Herabstufung des Reifegrades nach der Evaluation oder durch eine Verbesserung des Reifegrades infolge des Projektdurchlaufs und dem Abschluss des Einsatzes.

6.2.3 Datenqualitätsdimensionen

Wie in Abbildung 37 zu sehen ist, stellen die Datenqualitätsdimensionen den Konnex zwischen der theoretischen Prozessebene und der empirischen Reifegradebene her. Sie sind die wissenschaftlich theoretischen Mindestanforderungen, die in der Bewertung des Prozesses durch die Reifegrade nicht vergessen werden dürfen.

[445] Quelle: Eigene Darstellung

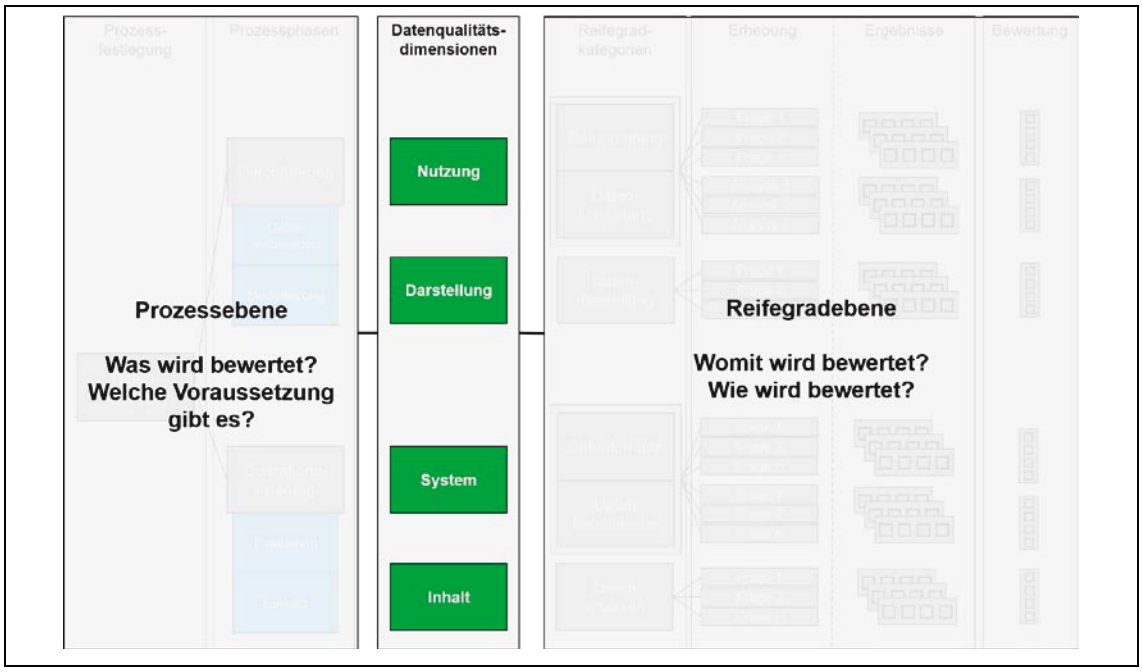

Abbildung 37: Datenqualitätsebene im Reifegradmodell[446]

Verwendet werden die Bezeichnungen der DGIQ bzw. die Anpassungen aus Abschnitt 3.4. Die Datenqualitätskategorien sind den zusammengefassten CRISP-DM Prozessphasen zuordenbar. Dabei wird die Durchführung durch die Kategorien Nutzung und Darstellung abgebildet und die Operationalisierung durch die Kategorien System und Inhalt.

Kategorie Nutzung

Diese Kategorie wird auch als zweckabhängige Kategorie bezeichnet und umfasst Datenqualitätsdimensionen, die die Daten auf deren Eignung zum Einsatz in der datenanalytischen Anwendung überprüft. Die Datenqualitätsdimensionen decken die Anforderungen der Durchführungsphase ab.[447]

Vollständigkeit:

Die Vollständigkeit bewertet die verlustfreie Erfassung, Übertragung und Transformation von Daten. Der Grund für fehlende Daten kann in der Datenerfassung liegen und muss technisch oder organisatorisch entfernt werden. Die Bewertung erfolgt in der Kategorie Datenumfang. Handlungsempfehlungen können jedoch für die Kategorien der Datenerfassung und der Datenbereitstellung abgeleitet werden, da es einen Zusammenhang zwischen diesen Kategorien gibt (siehe Abbildung 30).

Die Probleme von fehlenden Einträgen wirken sich auf die anwendbaren Algorithmen aus. Manche verzeihen fehlende Einträge, während andere nur vollständige Datensätze erlauben.[448] Entscheidungsbaumalgorithmen sind relativ robust gegenüber fehlender

[446] Quelle: Eigene Darstellung
[447] Vgl. Apel, D. et al. (2015), S. 24 f.
[448] Vgl. Kantardzic, M. (2011), S. 14.

Werte.[449] Regellernende Algorithmen gehören wie die Entscheidungsbäume zu den induktiven Lernverfahren und sind daher auch gegen fehlende Werte robust.[450] Künstliche neuronale Netze[451] und Fuzzyalgorithmen[452] sind weitere Vertreter von Algorithmen die sich gegen fehlende Werte robust erweisen. Selbstorganisierende Karten (self-organizing maps SOM) hingegen sind gegenüber fehlender Werte weniger tolerant.[453] Da es sich um strukturabbildendes Verfahren handelt, zu denen auch die Clusteralgorithmen gehören, sind dieses ebenfalls anfällig gegenüber fehlender Werte. Der Grund liegt im inhärenten Ziel der Methoden. Die Strukturentdeckung wird durch fehlende Werte verzerrt und spiegelt unter Umständen nicht die Realität wider.

Eine Möglichkeit mit fehlenden Werten umzugehen, ist es die entsprechenden Datensätze zu löschen. Wenn zu viele Datensätze betroffen sind, verringert dies die Aussagekraft der Datenbasis. Dieses Problem vergrößert sich mit einer großen Anzahl von Attributen.[454] Die in Abbildung 38 dargestellte Tabelle soll den Umstand verdeutlichen. Sollte jede Zeile mit einem fehlenden Wert gelöscht werden, bleiben von den ursprünglich elf Einträgen nur vier übrig.

ID	Attribut 1	Attribut 2	Attribut 3	Attribut 4	Attribut 5
1	7,5	rot	15,3	schwer	5
2	5,5	rot	12,1	schwer	3
~~3~~	~~6,3~~	~~blau~~		~~leicht~~	~~3~~
~~4~~		~~rot~~	~~19,0~~	~~mittel~~	~~1~~
~~5~~	~~6,4~~	~~blau~~	~~15,1~~		~~6~~
6	3,1	grün	23,0	schwer	6
~~7~~		~~blau~~	~~14,0~~	~~leicht~~	~~0~~
~~8~~	~~2,0~~		~~19,3~~		~~0~~
~~9~~	~~0,1~~	~~rot~~		~~mittel~~	
10	0,5	grün	11,1	mittel	2
~~11~~	~~1,3~~	~~gelb~~	~~5,5~~		~~5~~

Abbildung 38: Auswirkung fehlender Werte[455]

Bei zeitlich sequenziellen Daten ist die Löschung von Einträgen noch gravierender, da die Kontinuität einiger Attribute unterbrochen wird. Um Löschungen zu vermeiden, ist es möglich, fehlende Werte aufzufüllen. Dabei stellt sich die Frage, wie die Daten aufzufüllen sind.[456]

[449] Vgl. Cios, K. J. et al. (2007), S. 382.; Kantardzic, M. (2011), S. 189.; Kuhn, M.; Johnson, K. (2013), S. 380.
[450] Vgl. Cios, K. J. et al. (2007), S. 382 & 393.
[451] Vgl. Kantardzic, M. (2011), S. 201.
[452] Vgl. Kantardzic, M. (2011), S. 445.
[453] Vgl. Kantardzic, M. (2011), S. 230.
[454] Vgl. Emran, N. A. (2015), S. 119.
[455] Quelle: Eigene Darstellung
[456] Vgl. Petersohn, H. (2005), S. 62.; Cios, K. J. et al. (2007), S. 41 f.; Kantardzic, M. (2011), S. 36 f.

Eine Möglichkeit bei fehlenden nominalen Werten ist Verwendung eines Standardeintrags. Dadurch ist es möglich ein potenzielles Muster im Fehlen eines Wertes zu entdecken. Es wird dabei von der „informative missingness“ also dem *informationsbehafteten Fehlen* gesprochen.[457] Dieser Umstand kann in der Schwachstellenanalyse von Bedeutung sein, da fehlende Werte Hinweise auf Probleme bei Sensormessungen oder bei der Datenübertragung hinweisen.

Relevanz:

Die Relevanz unterscheidet sich von den vorherigen Dimensionen dadurch, dass ein Datensatz im angemessenen Umfang und vollständig vorliegen kann, der Inhalt jedoch für die Zielerreichung oder die Lösung der datenanalytischen Problemstellung irrelevant ist. Die Gründe von Irrelevanz von Daten können vielfältig sein. Die Daten stammen aus einem Umfeld, welches nicht mit der Zielstellung in Verbindung gebracht werden kann. Das Informationsangebot entspricht dabei nicht dem Informationsbedarf.[458] Einzelne Attribute können für die Lösung eines Problems unbrauchbar sein, da sie nur dieselben Einträge haben. Somit haben sie keinen Mehrwert. Oder die Werte korrelieren nicht mit der Zielvariablen und haben somit keinen Einfluss auf die Problemlösung. Mit Methoden der Datenreduktion können irrelevante Daten aus der Menge entfernt werden, um die Rechenzeit zu reduzieren und ungewolltes Rauschen zu verhindern.

Um relevante Attribute zu bewerten sind statistische Werte wie die Varianz oder die Spannweite für eine grobe Aussage geeignet. Eine Varianz von „0“ zeigt ein Attribut ohne Wertänderungen an. Um die Relevanz gegenüber eines Zielattributes zu bestimmen, sind Methoden der Informationstheorie, wie die Entropie oder Information Gain heranzuziehen.[459]

KUHN gibt für die Relevanz von Attributen konkrete Empfehlungen. Formel **6-1** bestimmt das Verhältnis des am häufigsten vorkommenden Wertes und des am zweithäufigsten vorkommenden Werts eines Attributes. Ist dieses Verhältnis zu groß, als Richtwert wird 20 gegeben, ist das Attribut sehr wahrscheinlich nicht für die weitere Verwendung geeignet. Das Risiko besteht darin, dass das Ungleichgewicht zwischen den Wertausprägungen Modelle in der Objektivität zu stark beeinflusst.[460]

$wert_{top}$…häufigster Wert
$wert_{top-1}$…zweithäufigster Wert

$$rel = \frac{|wert_{top}|}{|wert_{top-1}|} \qquad 6\text{-}1$$

Für eine datengestützte Schwachstellenanalyse sollten irrelevante Attribute entfernt werden, da diese zu Scheinzusammenhängen führen können. Solche Attribute sind zum Beispiel jene, die keine unterschiedlichen Werte haben. Somit würde der konstante Wert eines Attributes immer als eine Mitursache für eine Schwachstelle identifiziert werden.

[457] Vgl. Kuhn, M.; Johnson, K. (2013), S. 41.
[458] Vgl. Apel, D. et al. (2015), S. 27.
[459] Vgl. Kuhn, M.; Johnson, K. (2013), S. 377 ff.
[460] Vgl. Kuhn, M.; Johnson, K. (2013), S. 44 f.

Angemessener Umfang:

Nachdem unvollständige Einträge und irrelevante Attribute entfernt wurden muss der Umfang weiter angemessen sein. Das bezieht sich auf die Anzahl der Einträge, die sich durch die Aufzeichnungsdauer ausdrückt und die Anzahl der Attribute. Je komplexer die Analyse, desto länger sollte die Aufzeichnungsdauer sein, um ausreichend Daten erfasst zu haben.

Bei Analysen, in welchen die zeitliche Veränderung eine Rolle spielt muss der Aufzeichnungshorizont berücksichtigt werden. Fragestellungen, die durch saisonale Schwankungen beeinflusst werden sollten, so z. B. einen Zeithorizont von zwei bis drei Jahren umfassen.[461]

In Anwendungsfällen der Vorhersage und Klassifikation, also klassisches überwachtes Lernen, ergibt sich die Anforderung an den gemessenen Umfang aus der Verteilung der Daten. Die Data-Mining Verfahren sagen eine Zielvariable aufgrund einer Vielzahl von unabhängigen Attributen voraus. Je mehr Attribute vorhanden sind, desto mehr Datensätze sind nötig, um die Muster zu erkennen. Ansonsten spricht man von einem *imbalanced dataset*. Eine konkrete Aussage über die Häufigkeit kann nicht gemacht werden, es gibt nur Richtwerte.[462]

Relevant ist diese Dimension vor allem für die Modellierung, da die Ergebnisse eines Algorithmus damit stark zusammenhängen. Die Reifegradkategorie Datenumfang muss die Inhalte dieser Datenqualitätsdimension messen. Zu berücksichtigen ist die Dimension auch in der Datenaufbereitung, da hier die Datenbasis so bearbeitet werden muss, dass die negativen Auswirkungen sich nicht auswirken.

Aktualität:

Die Definition der Aktualität soll die Zeitstempelkonsistenz betrachten. Die Anwendung auf Prozess-, Betriebs-, und Maschinendaten ist nur so zielführend. Die Auslastung eines Tages wird sich genauso wie der Abnutzungsvorrat oder der Messwert eines Sensors verändern. Für die Schwachstellenanalyse hat die Aktualität in der neuen Definition besondere Bedeutung, um die Ursache-Wirkungsrichtung nicht zu verfälschen.

Die Aktualität wird im Reifegradmodell verwendet, um die Qualität des Zeitstempels zu bewerten. Wenn Meldungen verspätet und dann rückdatiert in das System eingegeben werden, ist die Qualität dieser Dimension negativ beeinflusst. Der Zeitstempel ist nicht mehr vertrauenswürdig, da die Rückdatierung sehr wahrscheinlich nicht sehr genau erfolgt. Des Weiteren fehlt der Eintrag unter Umständen zur Zeit der Analyse, wenn er benötigt wird.[463]

Wertschöpfung:

Die Wertschöpfung wird im Reifegradmodell nicht bewertet. Die Daten, bzw. deren hohe Qualität, sind wertschöpfend, wenn sie einen Mehrwert zum Unternehmen beitragen.

[461] Vgl. Dippold, R. et al. (2005), S. 243.
[462] Vgl. Kuhn, M.; Johnson, K. (2013), S. 44 f.
[463] Vgl. Woodall, P. et al. (2015), S. 330 f.

Der Mehrwert zeigt sich jedoch erst nach der Umsetzung eines Projektes und hier ist die Feststellung, welchen Beitrag die höhere Datenqualität hatte schwierig bis unmöglich. Die Literatur bietet im Gegensatz zur positiven Auswirkung von guter Datenqualität zahlreiche Beispiele wie sich schlechte Datenqualität negativ auf das Unternehmen auswirkt. Dabei sind die Beispiele hauptsächlich aus der Finanzwirtschaft, dem Einkaufs- oder Vertriebswesen sowie dem Kundenmarketing.[464] Für industrielle Anwendungen sind solche Beispiele nicht zu finden.

Kategorie Darstellung

Die Darstellungskategorie mit ihren Datenqualitätsdimensionen wird der Durchführungsphase zugeordnet, da diese auf die Möglichkeiten der Analyse Einfluss haben. So kann eine uneinheitliche Darstellung die Ergebnisse eines Algorithmus verzerren, da dieser ohne ein intuitives Vorwissen eines Menschen agiert. Wenn Daten sich widersprechen, oder ein Wert unterschiedlich dargestellt ist und trotzdem zum selben Ergebnis führen soll, kann die Mathematik eines Algorithmus damit nicht umgehen. Daher wurde diese Kategorie für die Datenaufbereitung und die Modellierung als wesentlich eingestuft.

Einheitliche Darstellung:

Die einheitliche Darstellung ist ein wesentlicher Faktor, um Daten vergleichbar zu machen und standardisiert auswerten zu können. Ein System muss daher so ausgelegt werden, dass eine einheitliche Darstellung unterstützt oder sogar garantiert wird. Eine Möglichkeit ist die Verwendung von vorgegebenen Codierungssystemen anstatt von Fließtexten bei Rückmeldungen und Aufzeichnungen.[465] Sind Daten nicht einheitlich dargestellt verringert sich der Umfang an aussagekräftigen Daten, da unterschiedliche Darstellungsformen die Häufigkeit einer Ausprägung verringern kann. In diesem Zusammenhang steht die Datenqualitätsdimension des angemessenen Umfangs eng in Abhängigkeit zur einheitlichen Darstellung. Die Möglichkeiten der Modellierung sind folglich davon abhängig. Die Qualität des Modellergebnisses und der möglichen Zielerreichung ist in anderer Weise von dieser Dimension abhängig. Wenn das Zielattribut eines überwachten Lernverfahrens unterschiedlich abgebildet wird, wird der Algorithmus versuchen unterschiedliche Muster in den Daten zu finden, um die uneinheitlichen Darstellungen eines Zielattributes vorherzusagen. Es kann zu Scheinzusammenhängen und einer geringeren Modellgüte kommen.

Für eine Schwachstellenanalyse ist die einheitliche Darstellung durch ein Codierungssystem wichtig. Dadurch können Auswertungen standardisiert durchgeführt werden und Mitarbeiter oder automatische Systeme melden auf diese Codes Probleme und Aktivitäten zurück.

Eindeutige Auslegbarkeit:

Die Grenze zur einheitlichen Darstellung ist fließend. Während die einheitliche Darstellung verhindern soll, dass ein Wert auf zwei unterschiedliche Weisen abgebildet

[464] Vgl. Apel, D. et al. (2015), S. 38 f.
[465] Vgl. Heidel, R. et al. (2017), S. 22.

wird, soll die eindeutige Auslegbarkeit bewerten, ob ein Wert auf unterschiedliche Weisen interpretiert werden kann.

Sie darf in der automatischen Auswertung mittels Data-Mining nicht vernachlässigt werden, da sich hier die Frage des Skalenniveaus stellt. Ist zum Beispiel eine physikalische Größe, wie das Gewicht, nur nominal angegeben und diese Angabe ist mit Zahlen codiert „1 – 2 – 3 – 4“, so kann das Skalenniveau als metrisch interpretiert werden. Folglich würde der Algorithmus nicht geeignete mathematische Operationen anwenden und falsche Ergebnisse liefern. Besser wäre eine textuelle Codierung wie „leicht – mittel – schwer – überschwer“.

Die durchgehende horizontale und vertikale Integration von Systemen rückt die eindeutige Auslegbarkeit in den Fokus. Bei der Übertragung der Daten sollten keine Unstimmigkeiten entstehen, weder entlang des Prozesses noch durch die Hierarchieebenen. Metadaten helfen dabei die korrekten Interpretationen vorzunehmen. In der Regel obliegt eine Interpretation jedoch Experten, da Computersysteme dazu noch nicht zuverlässig in der Lage sind.[466] Dieser Umstand ist für die Schwachstellenanalyse ebenfalls wichtig, da Daten auch von anderen Quellen herangezogen werden, um Ursachen und Zusammenhänge zu finden.

Kategorie System

Die Datenqualitätsdimensionen der Systemkategorie bewerten das Informationssystem, welches die Daten erfasst, überträgt, speichert und zur Verfügung stellt. Die Kategorie hat großes Gewicht für eine Operationalisierung der Ergebnisse in einem durchgehenden Datenanalyseprojekt. Wenn die Daten nicht einfach erfassbar, schnell übertragbar und transformierbar sind, kann ein erstelltes Modell nicht in den alltäglichen Echtzeit-Prozessablauf implementiert werden.

Von der Durchführungsphase wird diese Kategorie abgegrenzt, da es in der Durchführung darum geht, ob die vorliegenden Daten ausreichend gut für eine Analyse sind. In der Operationalisierungsphase geht es darum, wie die Daten zur Verfügung gestellt werden können und ob dies in einem automatischen Ablauf möglich ist. Daher ist die Systemkategorie der Operationalisierungsphase zugeordnet.

Zugänglichkeit:

Die Zugänglichkeit der Daten ist ein wichtiger Indikator, ob sie für die Analysen verwendet werden können. Für die einmalige Durchführung der Analyse können auch schwer zugängliche Daten aufbereitet werden. Der Fokus, der auf dieser Dimension liegt, ist jener der Automatisierbarkeit der Analysen. Die Daten müssen in einer Form gespeichert und auf eine Art erfasst werden, dass sie in einem automatisierten Workflow verwendet werden können.

Probleme entstehen, wenn Daten auf unterschiedliche Quellen verteilt sind. Sind IT-Systeme Drittsysteme auf die schwer oder unter hohen Kosten (wie z. B. Lizenzgebühren) Zugriff besteht, verringert dies ebenfalls die Zugänglichkeit für laufende automatische Analysen. Ein weiteres Problem bei Drittsystemen sind proprietäre

[466] Vgl. Heidel, R. et al. (2017), S. 22 f.

Formate[467].[468] Diese von Standardsystemen auslesbar zu machen kann ebenfalls aufwändig und damit teuer sein. Konvertierungen über andere Systeme oder sonstige Prozesse senken die Rechengeschwindigkeit, die in der Echtzeitanalyse kritisch ist. Eigene Systeme und Formate die Daten in ein Data-Warehouse speichern sind eine Abhilfe. Dieses Data-Warehouse bildet für alle Anwendungen den Singe Point of Truth. Neben proprietären Systemen sind auch veraltete Systeme problematisch. Diese können unter Umständen nicht an neue echtzeitfähige IT-Strukturen angebunden werden. Ein weiterer Grund für eine schlechte Datenzugänglichkeit kann das Outsourcing von Dienstleistungen sein. Wenn die Datenrückmeldung nicht direkt in das eigene System erfolgt ist es schwer die korrekte Eingabe zu garantieren, bzw. diese überhaupt zu erhalten.[469]

Ein spezieller Konnex zur Schwachstellenanalyse kann nicht hergestellt werden, da die Zugänglichkeit und die resultierenden Probleme falls diese schlecht sind, für alle Analysearten Allgemeingültigkeit haben.

Bearbeitbarkeit:

Die Bearbeitbarkeit beschreibt, wie leicht die Daten veränderbar sind. Im Rahmen von datenanalytischen Projekten stellt sich die Frage, ob die Daten im vorliegenden Format weiterverarbeitet werden können. Speziell in der Operationalisierung der Aufgabe kann dies zu einem Problem führen, da die Datentransformation schnell erfolgen muss. Eine einfache Darstellung in Standardformaten, die leicht auslesbar und welche die Daten in einem Standard-Skalenformat abbilden, sind ideal. Problematisch gestalten sich Bild- und Videodateien. Diese müssen erst interpretiert werden, bevor sie weiterverarbeitet werden können. Wenn die Daten in einem Vorverarbeitungsschritt auf die wesentlichen Merkmale reduziert werden, ist das wiederum kein Problem für die Weiterverarbeitung.[470] Zum Beispiel Qualitätsmessungen mittels Videosystemen. Das Ergebnis ob die Qualität in Ordnung ist, wird in einem Logfile abgelegt. Dieses Logfile kann weiterverarbeitet werden.

Für die Schwachstellenanalyse gilt das Gleiche wie für die Zugänglichkeit. Die Datenqualitätsdimension hat keine nur auf die Schwachstellenanalyse definierbare Auswirkung, sondern ist für Analysen jeder Art von Wichtigkeit.

Kategorie Inhalt

Der Inhalt der Daten beschreibt generische Metainformationen über die Daten. Hier geht es nicht darum wie Daten codiert sind, ob die Daten ausreichend sind oder deren Bedeutung interpretiert werden kann. Vielmehr beschreiben die Dimensionen dieser Kategorie inhärente Eigenschaften der Daten, die Unabhängig von der Nutzung sind und die sie immer haben. Deshalb ist diese Datenqualitätskategorie der Operationalisierung mit den generischen Handlungen und Zielen zugeordnet.

[467] Proprietäre Dateiformate: Nicht-Standardformate die von anderen Programmen nicht geöffnet werden können. Vgl. Matzer, M.; Lohse, H. (2007), S. 137.
[468] Vgl. Biedermann, H. et al. (2017), S. 30.
[469] Vgl. Woodall, P. et al. (2015), S. 327 ff.
[470] Vgl. Sonka, M. et al. (2015), S. 116 ff.

Glaubwürdigkeit:

Die Glaubwürdigkeit beschreibt wie sehr der Quelle der Datenerzeugung vertraut werden kann. Je höher die Automatisierung und technische Unterstützung der Datenerfassung ist, desto höher ist diese Datenqualitätsdimension einzustufen. Je stärker der Einfluss des Menschen bei der Datenbereitstellung ist, desto mehr muss die Reputation über längere Zeit erworben werden. Menschen geben aus unterschiedlichen Gründen, beabsichtigt oder unbeabsichtigt, fallweise Daten falsch ein[471].

Prozesse, die die Datenqualität sicherstellen gehören in diesem Zusammenhang zum Repertoire einer Organisation die Daten mit hohem Ansehen zur Verfügung stellen will. Zur Evaluierung und Bewertung des Modellergebnisses ist es wichtig, dass den Daten geglaubt werden kann.

Fehlerfreiheit:

Die Fehlerfreiheit wird am ehesten mit der Datenqualität in Verbindung gebracht. Diese Datenqualitätsdimension wird von der Art der Datenerfassung stark beeinflusst. Manuell erfasste Daten sind fehleranfällig, da Menschen unwissentlich oder wissentlich falsche Daten eingeben können. Eine automatische Aufzeichnung wirkt dem entgegen. Eine andere Möglichkeit ist die Kontrolle durch Vorgesetzte, die die Eingaben stichprobenartig kontrollieren und auf Fehler hinweisen, um den Qualitätsprozess weiterzutreiben. Die Schulung von Mitarbeitern ist eine zusätzliche Unterstützung. Wenn Mitarbeiter über die Probleme informiert werden, die falsche Daten verursachen und geschult werden was mit den Daten weiter gemacht wird, steigt die Hemmschwelle inkorrekte Daten absichtlich einzugeben und die Gewissenhaftigkeit Daten korrekt einzugeben.[472]

Eine Validierungsprüfung ist eine weitere Möglichkeit Fehler in Daten zu vermeiden. Automatische Prüfungen entdecken falsche Sensormessungen. Falsche Messungen können invalide Werte sein, die physikalisch oder aufgrund des Prozesses unmöglich sind. Solche Falschmessungen können durch logische Kontrollen entdeckt werden. Sensormessungen, die miteinander im Normalfall korrelieren und das plötzlich nicht mehr tun, deuten auf ein Problem hin. Um diese Fehler zu entdecken, bedarf es einer Kontrollinstanz für Daten, die sich mit diesen Problemen befassen.

Die Genauigkeit der Daten beeinflusst die Evaluierung der Ergebnisse des Projektes und die Erfolgswahrscheinlichkeit des gesamten Prozesses. Daher muss eine Aussage getroffen werden ob ein Datenanalyseprojekt mit korrekten Daten arbeitet. Die Durchführung der Analyse kann auch mit fehlerbehafteten Daten gemacht werden. Die Erfolgsaussichten sind jedoch geringer.

Die Beschreibung der Prozessphasen mit ihren Tätigkeiten und die Ausführungen zu den zugeordneten Datenqualitätsdimensionen geben den theoretischen Input der für die Ausgestaltung der empirischen Ebene. Zusätzlich werden die fallbeispielbezogenen Anforderungen berücksichtigt.

[471] Vgl. Haegemans, T. et al. (2018)
[472] Vgl. Woodall, P. et al. (2015), S. 331.

6.3 Empirische Ebene

Die empirische Ebene setzt sich auf Basis der in der theoretischen Ebene betrachteten Grundlagen mit der Bewertung im Reifegradmodell auseinander. Der Zusammenhang ist in Abbildung 32 dargestellt. Der Prozess der Datenanalyse wird in die zwei Metaprozessphasen aufgeteilt. Diese haben abhängig von den Vorgaben der Vorbereitungsphase Anforderungen. Diese ergeben die zum Teil durch die Datenqualitätsdimensionen, die ihnen zugeordnet sind. Die Dimensionen der Datenqualitätskategorien beschreiben ein Minimum an Anforderungen, die bei einer Reifegradbewertung betrachtet werden müssen. Die Reifegradkategorien fassen diesen theoretischen Input aus der Literatur auf und interpretieren und kombinieren ihn durch die Erkenntnisse aus den Anwendungen. Dadurch lassen sich die Reifegradkategorien beschreiben sowie die Erhebungsarten definieren und ausgestalten. In folgenden Abschnitten wird der Inhalt der Reifegradkategorien beschrieben. Diese Beschreibung erfolgt generisch und hat Allgemeingültigkeit für alle Analysearten. In einigen Fällen werden spezifische Beispiele gegeben, die sich auf die Schwachstellenanalyse beziehen.

6.3.1 Reifegradkategorien

Die Reifegradkategorien sind die Schnittstelle zwischen der Organisation und dem Anwendungsfall, anhand derer der Prozess der Datenanalyse bewertet werden soll. Sie wurden im Rahmen des DSR aus den Erfahrungen durch Projekte (Umwelt) und der Literatur definiert. Ihre Schnittstellenfunktion erfüllen sie, indem sie die abstrakten Begrifflichkeiten des CRSIP-DM und der Datenqualitätsbetrachtung auf anwendungsnahe Aspekte und Handlungsfelder der Praxis herunterbrechen.

Datenerfassung (Datenaufzeichnung)

Die Datenerfassung steht am Beginn eines datenanalytischen Prozesses. Wichtige Grundsteine für den Erfolg und den Arbeitsaufwand in einem Data-Mining Projekt werden hier gelegt. Darüber hinaus ist die Art der Datenerfassung ein wesentlicher Einflussfaktor für die spätere Implementierung des Ergebnisses eines Data-Mining Prozesses.

Die Reifegradkategorie bewertet, wie der Zustand eines Objektes, Prozesses oder der Umwelt in Form von Daten abstrahiert wird. Diese Erfassung der Daten kann in erster Linie manuell oder automatisch, digital oder nicht digital und regelmäßig oder unregelmäßig erfolgen. Abbildung 39 zeigt die Möglichkeiten der Datenerfassung als Würfel. Die höchste Reifegradausprägung ist das Würfelsegment 1-1-1, während jenes mit der geringsten Reife jenes mit den Koordinaten 2-2-2 wäre.

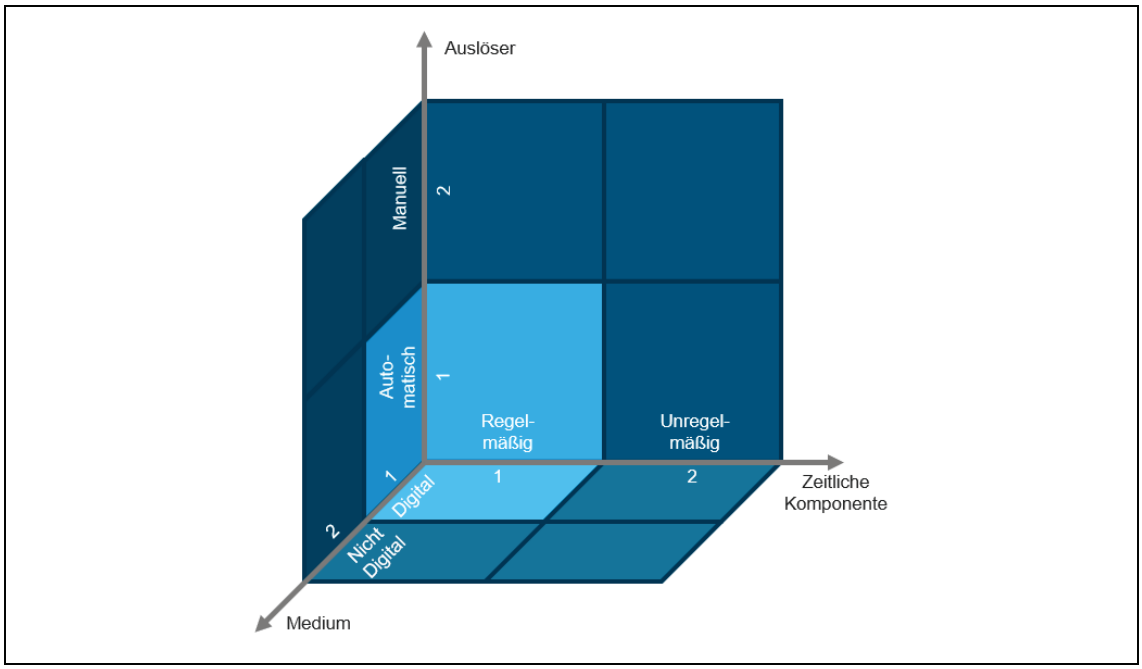

Abbildung 39: Datenerfassungswürfel[473]

Die manuelle Datenerfassung bedarf eines menschlichen Auslösers, während die automatische Datenerfassung unabhängig die Daten erfasst. Die digitale Datenerfassung zeichnet Daten mittels eines Sensors, einer Kamera, eines Mikrofons oder durch mobile Geräte oder Terminals auf. Die Maschinendatenerfassung[474] und die Betriebsdatenerfassung[475] sind wichtige Befähiger in der Smart Factory von Industrie 4.0. Die automatische und digitale Datenerfassung erlaubt es Daten in Echtzeit auszutauschen. Dieser echtzeitbasierte Datenaustausch ist ein wesentlicher Faktor am Weg zu Industrie 4.0.[476] Der Trigger bei der digitalen Aufzeichnung kann menschlich oder automatisch sein.

Digital-Manuell-regelmäßige Aufzeichnungen ergeben sich bei Rundgängen, die mit Mobile-Devices unterstützt werden oder wo der Mensch in regelmäßigen Rundgängen an Terminals Daten eingeben oder einen Status bestätigen muss.

Digital-automatisch-regelmäßige Aufzeichnungen sind Messungen und Zustandserfassungen mit Sensoren, über die MDE und BDE oder das Auslesen der

[473] Quelle: Eigene Darstellung

[474] MDE: Maschinendatenerfassung sammelt alle Daten die Maschinen betreffen. Die Dateneingabe kann automatisch, direkt aus der Maschinensteuerung über Bussysteme oder manuell – hier über Terminals oder andere Schnittstellen – erfolgen. Die Daten umfassen eine breite Spannweite von zustandsrelevanten Maschinendaten, bzw. prozessrelevanter Daten, die von der Maschine erzeugt werden. Die Daten können Ressourcen und Ressourcengruppen zugeordnet werden, um sie einer späteren automatischen Auswertung wie einer Schwachstellenanalyse zuführen zu können. (vgl. Kletti, J. (2015), S. 21 f.)

[475] BDE: Betriebsdatenerfassung zeichnet auftrags-, personen- und prozessbezogenen Daten auf. Die grobe Unterteilung erfolgt in Zeit- und Mengenaufzeichnungen. Bei der Aufzeichnung kann eine Unterscheidung in Gutstück und Ausschuss, sowie dessen Kategorien, erfolgen. Des Weiteren sind Abnutzungsvorräte und Materialverbräuche von Betriebsstoffen oder Hilfsstoffen ein möglicher Input. Über die Informationen des Planungssystems, können die Daten mit den Aufträgen abgeglichen und diesen zugeordnet werden. (vgl. Kletti, J. (2015), S. 21.)

[476] Vgl. Bogaschewsky, R.; Müller, H. (2016), S. 1.

SPS. Abhängig von der Messgröße und des Einsatzfalls ändert sich die Regelmäßigkeit, sprich die Abtastfrequenz.

Digital-manuell-unregelmäßig werden Daten aufgezeichnet, die bei bestimmten Ereignissen anfallen. Ein Beispiel ist die reaktive Instandhaltung, bei der der Instandhalter mit Mobile-Devices die Aktivitäten dokumentiert, bzw. der Maschinenbediener den Ausfall über ein Terminal meldet.

Digital-automatisch-unregelmäßig erfolgt die Datenaufzeichnung, wenn der automatische Trigger erst durch ein Ereignis ausgelöst wird. Diese Ereignisse können ein Maschinenstatus oder eine bestimmte Schwellwertänderung eines Sensors sein. Eine Sensoraufzeichnung basierend auf Schwellwertänderungen verringern die Datenmenge, erschweren jedoch die Zusammenführung von Daten. Es ergibt sich die Frage, wie die Lücken behandelt werden. Das Problem wurde in Abschnitt 6.2.3 „Vollständigkeit“ (Abbildung 38) erörtert.

Nicht digitale-automatische Aufzeichnungen existieren nicht. Nicht digitale-manuelle-regelmäßige und unregelmäßige Aufzeichnungen, unterscheiden sich von den digitalen, da die initiale Datenerfassung auf Papier erfolgt.

Die Datenqualitätskategorie „Inhalt“ gibt vor auf welche Umstände bei der Messung geachtet werden soll. „Glaubwürdigkeit“ und „Fehlerfreiheit“ sind bei digitalen Erfassungssystemen und jenen die automatisch ausgelöst werden am höchsten. Die Fehlerfreiheit wird durch automatische Routinen oder organisatorische Vorkehrungen, wie ein Rollenmanagement im Rahmen von Data Governance[477], erhöht.[478]

Datenhaltung (Datenbereitstellung und –transfer)

In dieser Reifegradkategorie wird untersucht, wie aufgezeichnete Daten für die Weiterverarbeitung bereitgestellt werden.

Schnittstellen und Medienbrüche führen bei der Datenübertragung zu Problemen und Leistungsverlusten. Zum einen besteht die Frage der Kompatibilität zwischen den Systemen und zum anderen geht es um Leistungseinbußen bei der Übertragungsgeschwindigkeit.[479] Eine Echtzeitübertragung der Daten vermeidet es, dass Daten im richtigen Moment fehlen oder bei einer nachträglichen Eingabe oder Synchronisation falsch in das System übertragen werden. Die Datenaufbereitung direkt an der Anlage reduziert die Datenmenge und erleichtert die Weiterverarbeitung in zentralen Systemen.[480] Ein spezielles Problem in diesem Zusammenhang bilden proprietäre Schnittstellen, die für Dritte nicht offen zugänglich sind.[481]

Die Datenübertragung und die Verbindung zwischen den Anlagen, sollte mit standardisierten Protokollen erfolgen, die problemlos von allen einsetzbar sind. Ein

[477] Data Governance: „Data Governance (Datensteuerung) umfasst in Summe die Menschen, Prozesse und Technologien, die zur Verwaltung und zum Schutz des Datenkapitals des Unternehmens benötigt werden, um allgemein verständliche, korrekte, vollständige, vertrauenswürdige, sichere und auffindbare Unternehmensdaten garantieren zu können.“ (Grosser, T. (2013), S. 2.)
[478] Vgl. Hastings, N. A. J. (2015), S. 229 ff.
[479] Vgl. Reder, L. et al. (2018), S. 13.
[480] Vgl. Schulmeyer, C. (2015), S. 310.
[481] Vgl. Heidel, R. et al. (2017), S. 78.

Beispiel dafür ist OPC-UA[482]. Des Weiteren erlaubt der Standard eine Übertragung zwischen neu installierten Komponenten von der Automatisierungsebene zur ERP-Ebene[483]

Ein Referenzarchitekturmodell ist ein Hinweis auf einen hohen Reifegrad in dieser Kategorie. Skalierbarkeit der Systeme, deren Interoperabilität, die transparente Vernetzung, eine integrierte und zuverlässige Datenverwaltung und die Fähigkeit zur Datenanalyse sind einige der Anforderungen, die an Industrie 4.0 Systeme gestellt werden.[484] Ein Data-Warehouse mit Data-Marts oder ein integrierter Data Lake erfüllen diese Anforderungen. Ein Referenzarchitekturmodell erlaubt neben der unternehmensinternen vertikalen Integration auch die unternehmensübergreifende horizontale Integration. Außerdem unterstützt es die Definition eines durchgängigen Datenmodells. Ein solches und relationale Datenbanken wirken sich positiv auf die Datencodierung und –darstellung aus. Datenbanken und Data-Warehouses haben in der Regel hochstrukturierte Daten.[485]

Eine nach wie vor häufig verwendete Form der Datenspeicherung, sind Exceltabellen (oder ähnliches),[486] welche in einer Ordnerstruktur auf einem Server abgelegt werden. Diese Daten können willkürlich kopiert und verschoben werden. Diese Lösungen erfüllen nicht die Anforderungen eines Single Point of Truth, im Gegensatz zu einem Data-Warehouse. Die Datenhaltung auf externen Speichermedien oder vor Ort bei Anlagen oder auf Papier entspricht nicht einem Industrie 4.0 Konzept.

Die Datenqualitätskategorie „System*“* bildet den Inhalt dieser Kategorie ab. Die Datenqualitätsdimension „Zugänglichkeit“ ist besonders geeignet eine Aussage zu treffen, wie die Datenhaltung und die Übertragung funktionieren.

Datenformate (Dateiformate)

In der Kategorie der Datenformate wird festgehalten, wie die Daten semantisch und syntaktisch abgebildet werden. Nicht offen zugängliche Definitionen, auch proprietäre Dateiformate, sind für Analysen mit Standardprogrammen nicht zugänglich. Die Formalisierung und Standardisierung zwischen den Daten ist ein Grundstein für die maschinengestützte Auswertung der Daten. Auch in diesem Fall helfen Referenzarchitekturmodelle und darin enthaltene Normen eine Durchgängigkeit und Kompatibilität zu gewährleisten.[487]

Datenformate von Bild-, Video- oder Tonaufzeichnungen sind in der Regel standardisiert und offen zugänglich. Eine schnelle Bearbeitung mit Standardprogrammen und die Extraktion der relevanten Daten ist jedoch aufwendig. Diese Extraktion sollte bereits bei der Aufzeichnung erfolgen, damit die relevanten Daten in strukturierter Form vorliegen.

[482] OPC-UA: Open Platform Communications Unified Architecture. Ein offen zugänglicher Schnittstellenstandard, der die Kommunikation zwischen Anlagen und deren Umfeld regelt. (Vgl. VDMA; Fraunhofer IOSB-INA (2017), S. 1.
[483] Vgl. Gartzen, T. et al. (2009), S. 52.; Heidel, R. et al. (2017), S. 48.
[484] Vgl. Heidel, R. et al. (2017), S. 4 ff.
[485] Vgl. Schulmeyer, C. (2015), S. 310.
[486] Vgl. Lueth, K. L. et al. (2016), S. 8.
[487] Vgl. Heidel, R. et al. (2017), S. 80 ff.

Das unterstützt die Echtzeitfähigkeit der Analyse indem die großen Datenmengen in kleinere Einheiten zerlegt werden.[488]

Abgelegt können die Ergebnisse in bekannten Datenformaten wie „csv“ oder Tabellenkalkulationsdateien wie „xlsx“ sein. Die Konvertierung dieser Datenformate ist wichtig, um einen einheitlichen Datenbestand zu schaffen. Durch die steigenden Datenmengen haben sich neue Datenformate etabliert, die es erlauben die Daten aufzuteilen. Dadurch ist es möglich diese auf mehreren Rechnern zu analysieren. Zu diesen Datenformaten zählen das HDF (Hierarchical Data Format) und HDFS (Hadoop Distributed File System). Datenformate mit Metainformationen sind gut verwendbar, da sie maschinenlesbare Daten über die Attribute enthalten, die für eine weitere automatisierte Interpretation wichtig sein können.[489]

Die Datenqualitätskategorie „System“ beurteilt die wichtigsten Faktoren dieser Reifegradkategorie. So sind proprietäre Formate schwer zugänglich und nicht bearbeitbar. Bilder sind zwar zugänglich, jedoch nicht bzw. schwer bearbeitbar.

Datendarstellung und –codierung

Diese Reifegradkategorie geht über die Betrachtung des Datenformats hinaus. Sie bewertet wie die Daten innerhalb einer Datei repräsentiert werden. Dabei spielt die Strukturierung der Daten, das Skalenniveau und die Standardisierung mit einem Codesystem eine Rolle.

Unstrukturierte Daten sind zum Beispiel Texte in natürlicher Sprache. Sie können von Computern nicht ohne vorgeschaltete Analysen interpretiert und analysiert werden. Daher sollten Daten in einer strukturierten Weise vorliegen. Die einfachste Möglichkeit sind dabei vorab definierte Standardcodes in nominalem Skalenniveau. Dabei ist darauf zu achten, dass die Darstellung der codierten Merkmale atomar erfolgt.[490] So kann eine Schadensinformation mit Standardcodes in der Form „2M05F10G01W35“ wiedergeben werden. Dabei ist der Störcode, der Schadensort und das Schadensmerkmal (Schadensbild) sowie die Ursache in diesen Code verschlüsselt. Besser wäre die Darstellung aufgeteilt in die einzelnen Attribute der Schadensmeldung, wie Tabelle 13 das zeigt.

Tabelle 13: Schadensinformation[491]

	Schadensort			
Störcode	**Maschine**	**Komponente**	**Schadensbild**	**Ursache**
2	M05 *Maschine 05*	F10 *Flansch*	G01 *gebrochen*	W35 *Werkstoff-fehler*

[488] Vgl. Schulmeyer, C. (2015), S. 310.
[489] Vgl. Kinz, A.; Bernerstätter, R. (2016), S. 66 f.; Taheri, M.; Mahdavi, A. (2018), S. 435.; Eickelmann, M. et al. (2019), S. 31.
[490] Vgl. Heidel, R. et al. (2017), S. 22 f.
[491] Quelle: Eigene Darstellung in Anlehnung an Pawellek, G. (2016), S. 124.

Eine solche Aufgliederung bringt den Vorteil, dass die einzelnen Codeteile einheitlich dargestellt werden und eindeutig analysierbar sind.

Metrische Skalenniveaus bilden die nächst höhere, bzw. die höchste Stufe beim Informationsgehalt von Skalenniveaus. Hier ist es möglich durch Klassenbildung das Skalenniveau, wenn für Algorithmen nötig, auf niederere Skalenniveaus zu transformieren. Der günstigste Fall ist das Vorhandensein von Metadaten, die vorhandene Attribute beschreiben. Diese maschinenlesbaren Daten erlauben es die dargestellten oder codierten Informationen weiter zu interpretieren und in Kontext zum Prozess zu setzten.[492] Metadaten sind in Markup Formaten wie HTML und XML enthalten oder in Datenbanktabellen, die wiederum auf Formate verzichten.

Die beiden Datenqualitätsdimensionen „einheitliche Darstellung“ und „eindeutige Auslegbarkeit“, der Datenqualitätskategorie „Darstellung“ definieren den Inhalt einer genaueren Erhebung des Inhalts dieser Reifegradkategorie gut. Dabei ist die „einheitliche Darstellung“ eine Voraussetzung, die erfüllt werden muss, um die „eindeutige Auslegbarkeit“ zu beurteilen. Erst wenn die Daten einheitlich dargestellt sind, kann darüber bestimmt werden, ob diese Darstellung eindeutig interpretierbar ist.[493]

Datenumfang

Diese Reifegradkategorie bestimmt die ausreichende Datenmenge. So benötigen Verfahren des maschinellen Lernens viele Daten, um allgemein gültige Aussagen treffen zu können. Für einfache Visualisierungen und manuell durchgeführte Analysen sind große Datenmengen weder nötig noch zielführend, da sie nicht handhabbar sind.

Um prognostische Analysen durchführen zu können sollte eine repräsentativ große Stichprobe der Daten verwendet werden. Für die tatsächliche Anzahl gibt es keine Vorgaben. Es sollten genügend sein, damit alle möglichen Muster für eine Zielvariable abgebildet werden. Im Rahmen der Datenaufbereitung kann mit statistischen Verfahren oder mit einfacher Visualisierung bestimmt werden, welche Variablen unnötig sind.[494]

Unüberwachte Algorithmen stellen geringere Anforderungen an die Daten. Es muss kein Abgleich mit einer Zielvariablen erfolgen. Muster oder Gruppen können auch in kleinen Datenmengen gefunden werden. Fehlende Werte sind ebenfalls nicht zwingend problematisch. Speziell bei nominal skalierten Daten, kann ein fehlender Wert mit einer Dummy Variable für das Fehlen markiert werden. Dadurch besteht die Möglichkeit, dass das Fehlen ein informatives Muster ergibt.

Die Datenqualitätsdimensionen der Datenqualitätskategorie „Nutzung“ haben für diese Kategorie besondere Bedeutung. Die Relevanz ist ein Mindestkriterium, welches für eine hohe Reife im Datenumfang erfüllt werden muss. Abhängig vom Einsatzzweck sind Daten relevanter oder nicht für die Zielerfüllung. Daten an falsch aufgezeichneten Aufnahmestellen sind irrelevant für das datenanalytische Problem.

Die Vollständigkeit ist die nächste Stufe, die erfüllt werden muss. Unvollständige Daten sind für viele Algorithmen schwer zu handhaben und auch jene die gegenüber Fehlstellen resilient sind, können Fehlinterpretationen nicht ausschließen. Nachdem die

[492] Vgl. Schulmeyer, C. (2015), S. 307.
[493] Vgl. Heidel, R. et al. (2017), S. 22.
[494] Vgl. Camm, J. D. et al. (2018), S. 424.

Vollständigkeit hergestellt wurde – durch Auffüllung oder Löschung – und irrelevante Daten entfernt wurden, kann mit der letzten anwendbaren Datenqualitätsdimension – dem angemessenen Umfang – eine Aussage über die Reife in dieser Kategorie getroffen werden. Der Datenumfang hat eine starke Auswirkung auf die datenanalytische Prozessreife.

Datenkonsistenz

Diese Reifegradkategorie bewertet ob die vorhandenen Daten die Realität abbilden. Der ausschlaggebende Faktor ist eine zeitliche Normierung der Daten. Die Grundvoraussetzung ist das Vorhandensein eines Zeitstempels, bzw. eine ausreichende Granularität. Abhängig vom datenanalytischen Ziel, ist es nötig unterschiedlich granulare und aktuelle Zeitstempel zu haben.

Der Zeitstempel muss mit der beabsichtigten Information, die hinter den Daten steht, übereinstimmen. Der zeitliche Bezug einer Tätigkeit, muss der Zeit der Durchführung entsprechen und nicht der Erfassung. Das gleiche gilt für alle anderen Schritte in einem Prozess, die messtechnisch oder anderwärtig erfasst werden. In vielen Fällen werden nur abschließende zeitliche Aufzeichnungen erfasst, die das Ende einer Tätigkeit zeigen, wobei das Eintreten einer Situation wichtiger ist.[495]

Diese Reifegradkategorie hängt indirekt mit den Kategorien „Datenerfassung“ und „Datenhaltung“ zusammen. Wenn die Datenerfassung technologisch unterstützt erfolgt und die Daten in Echtzeit übertragen werden, steigt die Wahrscheinlichkeit, dass der Zeitstempel in den Daten, mit der Zeit des erfassten Ereignisses übereinstimmt. Bei händischen Aufzeichnungen wird eine Zeitangabe in vielen Fällen vergessen und im Nachhinein abgeschätzt. Liegt dieser Fall vor, kann auf die Daten kein Algorithmus angewandt werden, der Ursache-Wirkungszusammenhänge findet, Zeitreihenanalyse durchführt oder Prognosen vornimmt. Wenn es keinen definierten Offset zwischen Erfassung und Speicherung der Daten gibt, ist es auch nicht möglich in der Datenaufbereitung die Konsistenz herzustellen indem der Offset berücksichtigt wird.

Um den Umstand der Synchronität zwischen Datenerfassung und Datenspeicherung abzubilden, wurde die Datenqualitätsdimension „Aktualität“ neu definiert. Diese bildet den Kern dieser Reifegradkategorie ab. Schlüsse werden aus der Art der Datenerfassung und Übertragung gezogen. Dabei fließen die technischen Möglichkeiten und die organisatorische Unterstützung ein. Des Weiteren findet die Granularität Eingang. Je höher diese ist, desto genauer können Ereignisse gereiht werden und Wirkungsketten, welche für Schwachstellen- und Ursachenanalysen wichtig sind, gebaut werden.

6.3.2 Erhebung

Die Einordnung des Unternehmens in die Reifegradstufen der Reifegradkategorien mittels Fragen und Datenqualitätsmetriken. Die Erhebung entspricht dem Datenverständnis aus Abbildung 33 bzw. Abbildung 34. Sie sollte mit den

[495] Vgl. Goel, S. et al. (2013), S. 164.

Domänenexperten erfolgen, die für die IT-Struktur, das Datenmanagement und für den untersuchten Prozess zuständig sind.

Folgend werden die Inhalte der Fragen und die Analysen der Reifegradkategorien beschrieben. Die Fragen selbst sind im Anhang 1A zu finden. Im Rahmen der Erhebung bei einer Organisation vor Ort, kann die Reihenfolge variieren. Zusätzlich können Fragen Aussagen für mehrere Reifegradkategorien geben. Der Fokus der Fragen und der Analysen liegt auf dem Anwendungsfall der Schwachstellenanalyse.

Generelle Fragen zu Einordnung des Anwendungsfalls

Diese Fragen klären nochmals das Ziel des Projektes ab. Des Weiteren sollen sie Aufschluss über die Organisation im Datenmanagement geben und festlegen, welche Datenquellen in weiterer Folge betrachtet werden. Für jede Quelle, müssen die Fragen der sechs Reifegradkategorien gestellt und die Gesamtheit der Daten aller Quellen bewertet werden.

Datenerfassung

Der Reifegrad in dieser Kategorie wird nur über Fragen erhoben. Sie haben die Dimensionen des Datenerfassungswürfels zum Inhalt und versuchen eine qualitative Einschätzung der Domänenexperten zu den Datenqualitätsdimensionen Glaubwürdigkeit und Fehlerfreiheit abzugreifen.

Datenbereitstellung

Die Fragen fokussieren das zentrale Thema der horizontalen und vertikalen Integration. Die unterschiedlichen Fragestellungen helfen die Reifegrade der Kategorie herauszuarbeiten und die Datenqualitätsdimension Zugänglichkeit zu bewerten. In dieser Kategorie besteht ein Zusammenhang zu den einleitenden Fragen des Fragebogens, da die Anzahl der Datenquellen die Gesamtausprägung beeinflusst. Viele Datenquellen führen zu vielen Schnittstellen, die speziell bei einer breiten horizontalen Integration Probleme verursachen können, wenn keine Standards vereinbart sind.[496]

Datenformate

Die zentrale Fragestellung, die hier beantwortet wird, ist die Art und die Vielfalt der Datenformate. Eine generelle Einschätzung „gut“ – „schlecht“ ist kann nicht gegeben werden. Proprietäre Formate sind eine Ausnahme; diese sind generell als niedrigster Reifegrad einzustufen. Die Kompatibilität untereinander und zum Analyseprogramm muss der Datenanalyst bewerten.

Datendarstellung und –codierung

Die Fragen des Fragenkatalogs können einerseits mit den Domänenexperten esprochen werden oder dienen andererseits als Checkliste bei der Analyse des Anfangsdatenbestandes. Die Formeln **6-2** bis **6-5** orientieren sich an Metriken in der

[496] Vgl. Hazen, B. T. et al. (2014), S. 78.

Literatur für die Glaubwürdigkeit, die hier auf die einheitliche Darstellung und die eindeutige Auslagbarkeit übertragen werden.[497]

Die Formeln 6-3 und 6-5 bewerten für jedes Attribut innerhalb einer Datenquelle, ob die Werte einheitlich dargestellt und eindeutig auslegbar sind. Was unter einheitlich dargestellt und eindeutig auslegbar zu verstehen ist, ist in Abschnitt 6.2.3 beschrieben.

$$n \dots Attribute \qquad einheitl.Darstel. = \left\lfloor \frac{\sum_{i=1}^{n} rate_{ed}(i)}{n} \right\rfloor \qquad 6\text{-}2$$

$$rate_{ed} = \begin{cases} 1 & ja \\ 0 & nein \end{cases} \qquad 6\text{-}3$$

$$n \dots Attribute \qquad eind.Auslegb. = \left\lfloor \frac{\sum_{i=1}^{n} rate_{ea}(i)}{n} \right\rfloor \qquad 6\text{-}4$$

$$rate_{ea} = \begin{cases} 1 & ja \\ 0 & nein \end{cases} \qquad 6\text{-}5$$

Ist das Attribut einheitlich bzw. eindeutig, erhält es den Wert „1“. Die Summe wird jeweils addiert, daraus der Mittelwert gebildet und abgerundet (siehe Formel 6-2 und 6-4). In einem zweiten Schritt kann das datenquellenübergreifend bewertet werden. Die Interpretation obliegt dem Datenanalysten. Ein Datenbestand mit weniger als 95% einheitlich dargestellter und eindeutig auslegbarer Attribute deutet auf eine geringe Reife hin. Die Vereinheitlichung der Daten ist mit viel manuellem Arbeitsaufwand verbunden.

Bewertung des Datenumfangs

Die Erhebung des Reifegrades zum Datenumfang erfolgt in Stufen. Zuerst werden die relevanten Attribute bestimmt. Dies erfolgt durch eine Visualisierung in Form eines Histogramms. Der Datenanalyst bestimmt daraufhin, ob die Wertvielfalt in einem Attribut relevant ist. Nicht relevante Attribute werden entfernt. Die Relevanz von Daten hängt zum Teil von der Aufgabenstellung ab. Für die Schwachstellenanalyse ist es wichtig die Daten an den richtigen Stellen zu erfassen. Darüber hinaus ist eine große Bandbreite an Einträgen nötig, um sinnvolle Auswertung durchführen zu können.

Danach folgt die Bewertung der Vollständigkeit nach den Formeln 6-6 und 6-7. Dafür wird mit einem Domänenexperten bestimmt, welche Werte als fehlend gewertet werden können.

$$n \dots Attribute \qquad V = \frac{\sum_{i=1}^{n} 1 - \frac{|fehlende\ Einträge|_i}{|Einträge|_i}}{n} \qquad 6\text{-}6$$

$$Vollst = \begin{cases} 4 & V \geq 0{,}99 \\ 3 & 0{,}99 > V \geq 0{,}95 \\ 2 & 0{,}95 > V \geq 0{,}90 \end{cases} \qquad 6\text{-}7$$

Datenbestände mit einer Vollständigkeit unter 0,95 werden mit „1“ bewertet. Hier muss versucht werden die fehlenden Werte aufzufüllen, die Datensätze zu löschen oder besonders schlechte Attribute zu entfernen. Welche Handlung gesetzt wird obliegt dem Datenanalysten. Sobald der Datenbestand bezüglich Vollständigkeit bereinigt wurde,

[497] Vgl. Frehe, V. et al. (2016), S. 149.

kann der angemessene Umfang bewertet werden. Dieser richtet sich nach der Anzahl der übrig gebliebenen Attribute. Weitere Erkenntnisse ergeben sich durch die Aufzeichnungsdauer der Daten.

Bewertung der Datenkonsistenz

Die Fragen behandeln vertiefend die Zeitstempelkonsistenz. Durch die Fragen und Antworten der Datenerfassung (Fragen 7, 8, und 11) sowie der Datenhaltung (Fragen 13,18,20 und 22) können Schlüsse gezogen werden, die auf den Reifegrad schließen lassen.

Die empirische Ebene definiert die Reifegradkategorien und gestaltet sie inhaltlich aus. Basierend darauf kann die Erhebungsmethodik abgeleitet werden, die den Inhalt abfragt. In der folgenden Bewertungsebene werden die Ergebnisse der Erhebung entsprechend der Beschreibung der Reifegradstufen der einzelnen Reifegradkategorien zugeordnet.

6.4 Bewertungsebene

Die Bewertungsebene setzt die Anforderungen und Erkenntnisse der empirischen Ebene um, um den Reifegrad in den Reifegradkategorien aufgrund des vorgegebenen Rahmens der Subebene „Erhebung“ einzuordnen.

6.4.1 Reifegradstufen

Im Reifegradmodell wurden vier Reifegradstufen definiert, anhand derer die Reifegradkategorien eingeordnet werden und der Prozess bewertet wird (siehe Kapitel 6). Die Reifegradstufen:

- Deskriptive Reife für den datenanalytischen Prozess → Reifegradstufe 1
- Diagnostische Reife für den datenanalytischen Prozess → Reifegradstufe 2
- Prädiktive Reife für den datenanalytischen Prozess → Reifegradstufe 3
- Präskriptive Reife für den datenanalytischen Prozess → Reifegradstufe 4

werden im folgenden Abschnitt beschrieben.

Reifegradstufe 1 – Deskriptive Reife

Reifegradstufe 1 erfüllt nur die Voraussetzungen für deskriptive Analysen wie jene der deskriptiven Statistik. Diese stellt Daten mithilfe einfacher numerischer oder visueller Verfahren dar.[498] Der Aufwand dieser Analysen ist gering, da keine großen Datenmengen vorliegen und diese noch einfach manuell aufbereitet und ausgewertet werden können. Die Digitalisierung ist kaum umgesetzt bzw. es existieren keine Standards.

[498] Vgl. Mittag, H.-J. (2012), S. 5 f.

Datenerfassung – Reifegradstufe 1:

Die Datenerfassung erfolgt ohne Standards und Ziele. Diese ist daher lückenhaft und ist als unzuverlässig einzustufen. In Abbildung 39 ist Reifegradstufe 1 in den Quadranten ‚Manuel', ‚nicht Digital' und ‚regelmäßig' oder ‚unregelmäßig zu finden. Die fehlende Unterstützung durch mobile digitale Systeme wie Smartphones oder Tablets zeigt eindeutig das Fehlen eines Standards und dass der Prozess der Datenerfassung, wenn CMMI als Basis herangezogen werden würde, auf Ad Hoc Aktionen basiert. Diese Aufzeichnungsart würde als wenig Glaubwürdig eingestuft werden. Organisatorisch gibt es keine Mechanismen eine gute Datenaufzeichnung sicherzustellen.

Die Aufzeichnung basiert unter anderem auf Papier und ist für höhere Analysen nicht mehr zeitgemäß. [499]

Die Datenmenge, die auf diese Weise erfasst werden kann, ist gering. Die Auswertungen, die für eine Schwachstellenanalyse durchgeführt werden können, beschränken sich auf einfache Visualisierungen und einfache arithmetische und statistische Verfahren mit dem Ziel einer ABC-Analyse.

Datenbereitstellung und -transfer – Reifegradstufe 1:

Die Daten liegen ursprünglich in einer für Analyseprogramme unzugänglichen Form vor. Ein Beispiel sind Papieraufzeichnungen die erst verspätet in ein isoliertes System übertragen werden. Dadurch kann den Zeitstempeln nicht vertraut werden. Eine weitere Möglichkeit sind proprietäre System, wie an Steuersystemen in Leitständen oder direkt an Anlagen.

Die Daten müssen mit viel Zeitaufwand zu einem gemeinsamen Datenbestand vereinheitlicht werden. Eine sinnhafte Zusammenführung ist nicht immer garantiert, daher ist es unter Umständen notwendig die Auswertungen auf den einzelnen Systemen durchzuführen und die Ergebnisse so zu interpretieren.

Datenformate – Reifegradstufe 1:

Die Daten liegen in einem proprietären Format vor und sind daher nicht oder nur unter hohem Arbeitsaufwand in ein gängiges Datenformat konvertierbar. Dieser oder die nötige Rechenzeit machen die Daten in solchen Formaten für automatisierte Echtzeitanalysen oder häufig durchgeführte Analysen unbrauchbar. Weitere wenig geeignete Formate sind Bild- und Videodateien. Deren nachträglicher Bearbeitungsaufwand ist hoch. Zum einen müssen relevante Merkmale aus den Dateien extrahiert werden und zum anderen ist der Rechenaufwand hoch.

Datencodierung und -darstellung – Reifegradstufe 1:

Auf Reifegradstufe 1 liegen Daten in unstrukturierter Form vor. Dazu zählen schriftliche Aufzeichnungen in Fließtextform. Eine weitere Ausprägung sind unbekannte und nicht interpretierbare Codes. Unstrukturierte Formate können mit erheblichem Arbeitsaufwand in ein standardisiertes und vergleichbares Format gebracht werden. Durch den

[499] Vgl. Drechsler, D. (2018), S. 247.

Arbeitsaufwand hält sich das Ausmaß, in dem Datenmengen erzeugt werden, die nur für einfache deskriptive Analysen verwertbar sind, in den Grenzen.

Datenumfang – Reifegradstufe 1:

Es ist kein angemessener Umfang an Daten gegeben. Die Gründe dafür können vielfältig sein. Zum einen sind die Daten unvollständig und daher kann keine durchgehende Datenbasis gebildet werden, da Datensätze oder Attribute gelöscht werden müssen. Zum anderen können die aufgezeichneten Daten für die Fragestellung die datenanalytisch gelöst werden geklärt werden soll, irrelevant sein. Irrelevante Daten sind z. B. Daten, die an falschen Messpunkten aufgezeichnet werden oder auf zu hoher Aggregationsstufe des Objektes oder des Prozesses erfasst werden. Die Datenmenge ist so gering, dass sie ohne Probleme in Tabellenkalkulationsprogrammen gehandhabt werden kann.

Datenkonsistenz – Reifegradstufe 1:

Der Fokus dieser Kategorie ist der einheitliche Zeitbezug der Daten. Auf Reifegradstufe 1 sind zeitliche Aufzeichnungen entweder nicht vorhanden oder sie sind besonders schlecht. Besonders schlechte Zeitstempel liegen vor, wenn die Datenerfassung und die Datenübertragung und -speicherung asynchron verlaufen und diese manuell erfolgt. Beispiele sind Aufzeichnungen von durchgeführten Arbeiten, die erst am Schichtende oder Ende der Woche aufgezeichnet oder übertragen werden. Weitere schlechte Zeitstempel sind jene, die nicht mit der zu untersuchenden Aktion in Verbindung gebracht werden können. Beispiele sind Daten einer Produktqualitätsmessung, die für Abnutzungsberechnungen bei Anlagen verwendet werden sollen. Wenn der zeitliche Bezug jedoch nicht den Produktionszeitpunkt wiedergibt, sondern den Zeitpunkt der Qualitätsmessung, ist dieser für die Anwendung unbrauchbar.

Reifegradstufe 2 – Diagnostische Reife

Die Analysen im Rahmen der diagnostischen Reife ermöglichen es Gründe für Ereignisse zu finden, bzw. schneller zu finden, da sie Muster und Zusammenhänge aufzeigen, die ohne die Analysen nicht ersichtlich gewesen wären.

Zum Einsatz kommen in erster Linie explorative Analysen der Statistik, die datengesteuert nach Mustern sucht.[500] Strukturprüfende Verfahren wie jene des unüberwachten Lernens sind diesen zugeordnet.

Datenerfassung – Reifegradstufe 2:

Die Datenerfassung hat in jedem Fall digital zu erfolgen, um einen Reifegrad von zwei oder höher zu erreichen. Der Trigger der Datenerfassung ist in der Regel manuell und unregelmäßig. Durch die Unterstützung technischer Systeme steigt die Glaubwürdigkeit der Datenerfassung an. Sie erlauben es Fehlerkorrekturmaßnahmen zu hinterlegen, um

[500] Vgl. Hedderich, J.; Sachs, L. (2016), S. 15.

offensichtlich falsche Eingaben zu vermeiden. Es werden Bluetooth, RFID, NFC oder Zigbee verwendet, um die Daten am Objekt zu erfassen.[501]

Die Datenmengen ermöglichen Korrelationsanalysen zur Entdeckung gegenseitiger Abhängigkeiten oder Clusteranalysen, um Zusammenhänge als Gruppen in den Daten wiederzufinden. Für die datengestützte Schwachstellenanalyse wird hier der Grundstein gelegt. Die Daten werden bereits digitalisiert erfasst, sodass eine spätere Analyse einfach zu automatisieren oder duplizieren ist. Um die manuelle Erfassung zu erleichtern, sollte eine Logik zwischen dem Prozess, den erzeugten Daten und den zu erfassenden Daten und Merkmalen bestehen. Zum Beispiel sollen bei Störungen an einem Elektromotor keine spezifischen Codes für Fehler und Ursachen von Hydraulikanlagen in der Eingabemaske aufscheinen.

Datenbereitstellung und -transfer – Reifegradstufe 2:

Die erfassten Daten werden in offenen, jedoch nicht vernetzten Systemen erfasst. Es handelt sich dabei um lokale Serversysteme. Es ist möglich Abfragen und Exporte vorzunehmen und diese, in weiterer Folge, mit jenen anderer lokaler Datenhaltungssysteme zu vereinheitlichen. Kompatibilitäts- und Schnittstellenprobleme können nicht ausgeschlossen werden. Der Arbeitsaufwand ist durch den manuellen Aufwand groß.[502]

Die Übertragung der Daten in das System erfolgt nicht in Echtzeit. Die Synchronisation mit den Erfassungssystemen erfolgt, wenn diese in dafür vorgesehenen Schnittstellenpunkten sind.

Datenformate – Reifegradstufe 2:

Auf dieser Reifegradstufe sind die Datenformate standardisiert und im Gegensatz zu Reifegrad 1 von den meisten Programmen uneingeschränkt lesbar. Beispiele sind Formate von Tabellenkalkulationsprogrammen (xls und xlsx), Textformate, PDFs oder XML. Die Zusammenführung zu einem einheitlichen Datenbestand ist, abhängig von der Menge und der Konstellation der Formate mit großem Zeitaufwand verbunden.

Datencodierung und -darstellung – Reifegradstufe 2:

Die Daten liegen in strukturierter Form, wie Standardcodes, vor. Das vorherrschende Skalenniveau ist nominal. Die Attribute sind nicht atomar eingetragen, was zu einem höheren Datenaufbereitungsaufwand führt, um die Kombinationen von Codes in Einzelattribute aufzuteilen. Die Vergleichbarkeit der Einträge erlaubt es standardisierte explorative Analysen durchzuführen. Datengestützte Schwachstellenanalysen sind möglich. Eine Grundvoraussetzung ist jedoch die eindeutige Auslegbarkeit der Codes. Dazu müssen diese auf den Zweck abgestimmt werden. Speziell in der Schwachstellenanalyse müssen klare Definitionen von Schadensbild und -ursache eingehalten werden.

[501] Vgl. Lee, G.-Y. et al. (2018), S. 988.
[502] Vgl. Dettke, C. et al. (2009), S. 121.

Datenumfang – Reifegradstufe 2:

Die Datenmenge in Reifegradstufe 2 ist in der Dimension der Datensätze (Einträge) und der Datenfelder (Attribute) größer als bei Reifegradstufe 1. Haupteinflussparameter ist die Aufzeichnungsdauer. Diese sollte wenigstens neun Monate betragen, um wesentliche Zusammenhänge, die sich unter Umständen im Jahresverlauf ergeben, erfassen zu können. Da explorative Analysen wie zum Beispiel die Clusteranalyse oder Assoziationsanalyse eingesetzt werden, sollte der Umfang der Attribute deren Einsatz rechtfertigten. Diese werden herangezogen, wenn nicht nur die Menge der Daten im Sinne der Datensätze groß, sondern auch deren Dimensionalität zu groß ist. Die Dimensionalität ist dann zu groß, wenn Strukturen und Zusammenhänge vom Menschen ohne Unterstützung nicht mehr gefunden bzw. analysiert werden können. Diese Grenze ist bereits bei drei Variablen erreicht.[503] Diese Grenze wird für den Datenumfang in Reifegradstufe „2“ gesetzt. Alles darunter ist nicht ausreichend für diagnostische Analysen. Die Vollständigkeit muss mit 0,5 laut Formel 6-7 bewertet werden.

Datenkonsistenz – Reifegradstufe 2:

Auf Reifegradstufe 2 ist ein Zeitstempel vorhanden, der einen Dateneintrag zeitlich einordnet. Die Zeiteinträge sind nur innerhalb einer Quelle gesichert vergleichbar. Das heißt innerhalb einer Quelle kann eine Reihung der Ereignisse vorgenommen werden. Datenquellenübergreifend kann die Konsistenz nicht garantiert werden. Mit gewissem Arbeitsaufwand in der Datenaufbereitung kann die Konsistenz jedoch hergestellt werden. Die Daten eignen sich dadurch für diagnostische Zwecke, da die Aktionen und Prozessschritte in eine korrekte Abfolge gebracht werden können.

Reifegradstufe 3 – Prognostische Reife

Reifegrad drei ist der Übergang zur vollständigen horizontalen und vertikalen Integration der IT-Systeme. Daten werden in einigen Fällen automatisch aufgezeichnet und Standards bei den Formaten und der Art der Repräsentation umgesetzt. Eine Analyse ist möglich, wenn auch die Zusammenführung aus unterschiedlichen Datenquellen mit erheblichem Datenaufbereitungsaufwand verbunden ist. Daten aus einer Quelle sind in sich als konsistent zu betrachten; datenquellenübergreifend ist die Konsistenz nicht garantiert. Des Weiteren ist eine Implementierung mit größerem technischem und organisatorischem Aufwand verbunden, da die Infrastruktur dafür noch nicht vorgesehen ist.

Datenerfassung – Reifegradstufe 3:

Die Datenerfassung sollte automatisch erfolgen. Sie wird jedoch unregelmäßig durch vordefinierte Trigger durchgeführt. Dadurch ist es schwieriger einen einheitlichen Datenbestand zu erzeugen. Darüber hinaus besteht die Gefahr sonstige wesentliche Zustände nicht aufzuzeichnen die zwischen den Triggerpunkten bei anderen Signalquellen passieren aber nicht den Trigger auslösen. Erfolgt die Aufzeichnung

[503] Vgl. Halford, G. S. et al. (2005), S. 75 f.

manuell, so muss diese digitalisiert mit einer technischen Unterstützung regelmäßig passieren. Dafür müssen organisatorische Maßnahmen getroffen werden.

Es gibt automatische Fehlerkorrekturmaßnahmen, die verhindern, dass falsche Werte aufgezeichnet werden können. Es handelt sich hierbei hauptsächlich um technische Systeme, die logische Schlüsse ziehen, um Fehler aufzudecken.

Datenbereitstellung und -transfer – Reifegradstufe 3:

Die Daten werden auf einem zentralen Datenbanksystem erfasst. Die Abfragen werden erleichtert, da Schnittstellenprobleme entfallen. Die Daten, die in dem Datenbanksystem erfasst werden sind bereichsübergreifend, jedoch nicht hierarchieübergreifend. Es handelt sich um die klassische horizontale Integration. Beispiele sind operative Daten der Instandhaltungsdurchführung, der Produktion und der Maschinenüberwachung.[504]

Die Übertragung vom Erfassungspunkt erfolgt in Echtzeit. Die Erfassungszeit ist somit die Speicherzeit. Daten stehen sofort für weitere Analysen zu Verfügung. Um große Datenmengen zu reduzieren, werden Vorarbeiten, wenn möglich und nötig direkt vor Ort bei der Messung durchgeführt. Diese Art der dezentralen Datenvorverarbeitung wird als Edge Computing bezeichnet. Der Vorteil ist, dass bei Messvorgängen, bei denen unstrukturierte Daten erzeugt werden, wie Bilder oder Videos, diese sofort auf die für die Anwendung relevanten Merkmale reduziert werden und dieser Datensatz in weiterer Folge für Analytics-Anwendungen in Kombination mit anderen Daten automatisch auswertbar ist.[505]

Datenformate – Reifegradstufe 3:

Die Datenformate sind in der Lage große Datenmengen zu speichern. Beispiele sind Formate mit strukturiertem Text wie csv. Die Kombination der Datenformate ist nicht hinderlich bei der Erstellung eines gemeinsamen Datenbestandes.

Datencodierung und -darstellung – Reifegradstufe 3:

Wesentliche Attribute sind metrisch skaliert. Dabei sind speziell die unabhängigen Variablen im Fokus auf denen eine Prognose von abhängigen Variablen durchgeführt werden. Ein entscheidender Vorteil von metrisch skalierten Werten ist, dass diese in nominal skalierte Werte transformiert werden können. Ein Augenmerk liegt auf der abhängigen Variablen. Die eindeutige Auslegbarkeit muss gegeben sein, damit nicht die in Abschnitt 6.2.3 beschrieben Probleme der Mustererkennung eintreten.

Datenumfang – Reifegradstufe 3:

Die Aufzeichnungsdauer in dieser Reifegradstufe muss wenigstens ein Jahr betragen. Die Vollständigkeit muss mit 0,5 laut Formel 6-7 bewertet werden und die restlichen Attribute müssen für den Prognoseprozess relevant sein. Die Empfehlungen aus **6-1** geben dafür Richtwerte. Das Augenmerk liegt auf der vorherzusagenden Variablen, dem Label. Dieses Attribut muss je Ausprägung, die vorhergesagt werden soll, mehrere Einträge haben. In der Literatur sind dazu keine Empfehlungen zu finden. Eine finale

[504] Vgl. Hastings, N. A. J. (2015), S. 25.
[505] Vgl. Staar, B. et al. (2017), S. 53 f.; Gatica, C. P.; Boschmann, A. (2019), S. 109

Einschätzung obliegt dem Data Scientist. Die Erfahrung aus Projekten zeigt, dass 15 je Ausprägung ein guter Wert ist, um eine angemessene Prognosegüte zu erhalten.

Datenkonsistenz – Reifegradstufe 3:

Die Zeitstempelkonsistenz ist über Datenquellen hinweg vergleichbar. Es existieren keine Offsets und der Zeitstempel erlaubt nicht nur eine gesicherte Reihung der Einträge, wie sie bei diagnostischen Zwecken eine Mindestanforderung ist, sondern er sichert auch den Abstand zwischen den Einträgen. Ein definierter zuverlässiger Abstand zwischen den Einträgen ist bei Prognosen von besonderer Bedeutung.

Reifegradstufe 4 – Präskriptive Reife

In Reifegrad vier ist die vollständige horizontale und vertikale Integration vollzogen. Ein durchgängiges Daten- und Informationsmanagement sichert Standards in der Datenaufzeichnung, den Formaten, der Darstellung sowie der Speicherung und Weitergabe. Um eine durchgehende Standardisierung über die Unternehmensgrenzen zu gewährleisten empfiehlt sich ein Referenzmodell einzusetzen.

Die Analyse setzt eine Kombination komplexer Algorithmen ein, um die Frage zu beantworten „Was soll geschehen?“. Wie bei der prädiktiven Analyse werden Vorhersagen auf der Grundlage aktueller Parameter und verfügbarer historischer Daten durchgeführt. Im Gegensatz dazu berücksichtigt die prädiktive Analyse jedoch Beziehungen, für die es keine historischen Daten gibt, die aber eine definierte Wahrscheinlichkeit haben, dass sie auftreten werden.[506] In der Wissenstreppe aus Abbildung 9 verhilft die präskriptive Analyse einem System richtig oder autonom zu handeln. Es wird dabei auch von Self-X Fähigkeiten eines Systems gesprochen.[507]

Datenerfassung – Reifegradstufe 4:

Die Datenerfassung erfolgt vollautomatisch in regelmäßigen Abständen. Die manuelle Eingabe von Daten ist nur mehr in Form von Bestätigungen von Werten nötig. Diese Werte werden aus einer Liste vorgegeben, die bereits nach den wahrscheinlichsten gefiltert und sortiert wurde, um die Eingabequalität zu verbessern. Damit ist eine maximale Glaubwürdigkeit und durch die Logik der Eingabe eine bestmögliche Fehlerfreiheit garantiert.

Datenbereitstellung und -transfer – Reifegradstufe 4:

Die Daten werden in einem Data-Warehouse gespeichert, die über Data-Marts diese für unterschiedliche Zwecke zur Verfügung stellen. Mittels OLAP werden einfache Datenaufbereitungen und Vorverarbeitungsschritte durchgeführt, deren Ergebnisse für die Analysen zur Verfügung gestellt werden. Zeitkritische Vorverarbeitungsschritte werden direkt an der Anlage durchgeführt. Sollten die großen Datenmengen, die auf dieser Ebene anfallen werden, in einen Data Lake gespeichert werden, so muss dieser durch die Funktionalitäten eines Data-Warehouses ausgelesen werden, damit die Daten in der nötigen strukturierten Form für die Analysen zur Verfügung stehen.

[506] Vgl. Sappelli, M. et al. (2017), S. 45 ff.
[507] Vgl. Heidel, R. et al. (2017), S. 16 f.

Datenformate – Reifegradstufe 4:

In Reifegrad „4“ sollten Datenformate nicht mehr nötig sein, da die Datenübertragung direkt über Schnittstellen zwischen den Datenbereitstellungsmedien und der Analysesoftware erfolgt. Sollte die Übertragung dennoch mittels Exports erfolgen, dann sind Datenformate zu wählen, die für Big Data geeignet sind. Ein Beispiel für ein solches Datenformat ist das Hierarchical Data Format (HDF)[508].

Datencodierung und -darstellung – Reifegradstufe 4:

Die Daten sind durchgehend metrisch skaliert oder als standardisierte Codes nominal dargestellt. Wichtig ist, dass diese Standardisierung der Codes durchgehend bei allen Datenquellen durchgezogen ist, um Interpretationsschwierigkeiten zu vermeiden. Des Weiteren sind die Codes atomar hinterlegt. Für die Datenquellen sind Metadaten enthalten, die automatische Interpretationen erlauben.

Datenumfang – Reifegradstufe 4:

Der Aufzeichnungsumfang beträgt wenigstens 1,5 Jahre. Die Daten sind vollständig erfasst. Bei den Datenquellen, die in der vertikalen Integration hierarchisch höhergestellt sind, wie Ziel- und Planungsdaten, ist darauf besonders zu achten. Die Relevanz ist durch eine genaue Zuordnung auf die jeweiligen Betrachtungseinheiten gegeben. Um komplexere Analysen und Simulationen zu ermöglichen ist die Attributanzahl mindestens im höheren zweistelligen Bereich.

Datenkonsistenz – Reifegradstufe 4:

Die Zeitstempelintegrität ist durch die Echtzeitübertragung und ein horizontal wie vertikal durchgängiges System garantiert.[509] In Kombination mit ID-Daten, wie Auftragsnummern oder Materialnummern, kann eine rückverfolgbare Konsistenz hergestellt werden. Diese ist für die Erstellung von komplexen Wirkungsnetzte z. B. im Sinne von Process-Mining nötig.[510] Für die Messung von Effizienzen werden von Aktivitäten Start- und Endzeitstempel erfasst.

6.4.2 Bewertung

Die Bewertung setzt die Reifegrade mit den Ergebnissen der Erhebungsmethoden in den Reifegradkategorien in Verbindung. Folgend wird diese Zuordnung beschrieben. Es wird explizit festgehalten, dass es sich um Mindestanforderungen handelt. Zusatzinformationen während der Reifegradbewertung können die Reifegradeinstufung beeinflussen. Die Zusammenführung zu einem Gesamtreifegrad in den Kategorien, wird unter der Annahme beschrieben, dass die Problemstellung eine Schwachstellenanalyse ist.

[508] Vgl. Taheri, M.; Mahdavi, A. (2018), S. 436.
[509] Vgl. Gartzen, T. et al. (2009), S. 50 f.
[510] Vgl. Van der Aalst, W. (2016), S. 30 ff.

Datenerfassung

Tabelle 14 zeigt die Einordnung der Antworten des Fragekatalogs die den Reifegraden eindeutig zugeordnet werden können. Die Restlichen müssen entsprechend der Reifegradbeschreibungen und des Kontextes zugeordnet werden. Dazu dient Tabelle 14 als Referenz.

Tabelle 14: Bewertung Datenerfassung[511]

Reifegrad	Glaubwürdigkeit	Fehlerfreiheit	Allgemein
1	7.b.i | 12.a		10.b
2	8.a	5.a | 11.b.i	6 | 10.d
3	7.a	5.b	10.c
4	12.d	11.b.ii | 27.a	7.a | 10.c

Die Einordung der weiteren Antworten von Frage 7.b, 9 und 10 hängt von der Quelle und den Anforderungen ab. Instandhaltungsdaten von Ausfällen müssen nicht zwingend vom Instandhaltungsmitarbeiter erfasst werden. Die Rückmeldung kann auch durch den Produktionsmitarbeiter an einem Terminal erfolgen. Frage 27 wird zwar in einer anderen Reifegradkategorie gestellt. Die Auswirkung ist jedoch auf die Fehlerfreiheit übertragbar.

Für die Zusammenführung zu einen Gesamtreifegrad in der Reifegradkategorie, sind die Ergebnisse aus der Spalte „Allgemein“ höher zu gewichten. So ist 12.a nicht der ausschlaggebende Faktor, wenn die Daten automatisch erfasst werden (10.c)

Datenbereitstellung

Tabelle 15 zeigt die eindeutigen Zuordnungen der Antworten zu den Reifegraden sortiert nach der zugeordneten Datenqualitätsdimension und darüberhinausgehenden Faktoren.

Tabelle 15: Bewertung Datenbereitstellung[512]

Reifegrad	Zugänglichkeit	Allgemein
1	14.b.ii | 18.a, i	20.a | 19.b
2	14.b.i | 18.c, d	13.b | 20.b
3	14.a | 18.e, f, h	13.a | 19.a | 21.b
4	18.f	15.a | 16.a & 17.a

Die Fragen 16 und 17 ergeben in Kombination Reifegrad 4, da hier die vollständige horizontale und vertikale Integration bewertet wird. Auf die allgemeine Einordnung wird die Antwort von Frage 22 nach sonstigen Problemen Einfluss haben.

[511] Quelle: Eigene Darstellung
[512] Quelle: Eigene Darstellung

In der Bildung des Gesamtreifegrades der Reifegradkategorie ist der Zugänglichkeit ein höheres Gewicht zu geben. Unzugängliche Daten und Systeme beeinflussen den Prozess ungleich stärker als z. B. das Fehlen eins expliziten Single Point of Truths.

Datenformate

Die vorgenommene Zuordnung in Tabelle 16 ist in dieser Reifegradkategorie relativ einfach durchzuführen. Es muss jedoch beachtet werden, dass eine ungünstige Kombination von Formaten bei der Schaffung eines gemeinsamen Datenbestands problematisch sein kann.

Tabelle 16: Bewertung Datenformate[513]

Reifegrad	Bearbeitbarkeit	Allgemein
1	23.l,23.i,23.j,23.k | 24	
2	23.a,23.c,23.d,23.e,23.f,23.g	Große Kombination der Formate
3	23.b,23.h	Ein einziges Format
4	23.m	

Es obliegt dem Datenanalysten festzulegen, welche Kombination aus den Datenformaten zu groß und ungünstig ist. Die Reifegrade drei und vier sind hingegen relativ einfach zu identifizieren.

Datendarstellung

Tabelle 17 kombiniert für die Reifegradbewertung die Ergebnisse von qualitativen und quantitativen Erhebungsmethoden.

Tabelle 17: Bewertung Datendarstellung[514]

Reifegrad	Einheitliche Darstellung	Eindeutige Auslegbarkeit	Allgemein
1	Formel 6-2 < 0,75	Formel 6-4 < 0,75	25.b.i | 28.b.ii | 29.d
2	Formel 6-2 ≥ 0,75	Formel 6-4 ≥ 0,75	25.b.iii | 28.b.i | 26.b.i | 26.b.iv
3	Formel 6-2 ≥ 0,95	Formel 6-4 ≥ 0,95	28.a | 29.a | 26.b.v
4	Formel 6-2 ≥ 0,99	Formel 6-4 ≥ 0,99	30.a | 31.a

Es ist außerdem nötig die Betrachtung quellenübergreifend vorzunehmen. Dazu eignet sich die einheitliche Darstellung, die beantwortet ob alle Attribute in verteilten Quellen durchgängig einheitlich dargestellt sind.

Der Gesamtreifegrad gewichtet die Fragen der Allgemein-Spalte am höchsten, danach die einheitliche Darstellung.

[513] Quelle: Eigene Darstellung
[514] Quelle: Eigene Darstellung

Datenumfang

Der Datenumfang wird in den Mindestanforderungen durch Tabelle 18 bewertet. Frage 36 ist bei überwachten Lernverfahren oder bei der spezifischen Suche nach Schwachstellen wichtig. Ansonsten ist die Relevanz durch den Datenanalysten zu bestimmen. Dazu wurden die beiden Extremausprägungen vorgegeben.

Tabelle 18: Bewertung Datenumfang[515]

Reifegrad	Relevanz	Vollständigkeit (Formel 6-6)	Angemessener Umfang	Allgemein
1	Frage 36 < 10 Nur ein Wert	V < 0,90	32.a, e | 34.a	4.c | 18.i
2	Frage 36 ≥ 10	V ≥ 0,90	32.c | 34.b	35.b.i
3	Frage 36 ≥ 20	V ≥ 0,95	32.d | 34.c	4.a | 35.b.iv
4	Frage 36 ≥ 50 Ausgeprägte Verteilung (Normal- oder Gleichverteilung)	V ≥ 0,99	32.f	

Es ist wichtig festzuhalten, dass der angemessene Umfang vor und nach der Bereinigung der irrelevanten und unvollständigen Werte betrachtet werden muss. Ist der Umfang vorab zu gering, ist eine Bereinigung nicht nötig, da eine solche den Umfang noch mehr verringern wird.

Bei der Zusammenführung der einzelnen Betrachtungsfelder, wird dem angemessenen Umfang das höchste Gewicht gegeben, danach den allgemeinen Punkten und abschließend der Vollständigkeit.

Datenkonsistenz

Tabelle 19 stellt für die Bewertung der Datenkonsistenz eine Mindestreferenz dar, die sich sehr stark auf die Zeitstempel konzentriert. Durch die Beantwortung in den anderen Reifegradkategorien, besonders in der Datenerfassung und der Datenbreitstellung, können zusätzliche korrigierende Erkenntnisse abgeleitet werden.

Tabelle 19: Bewertung Datenkonsistenz[516]

Reifegrad	Aktualität	Allgemein
1	7.b.i	37.a | 40.b.iv
2	38.b	37.b.iii | 37.b.ii | 40.b.iii
3	38.a	39.a | 39.b | 40.b.ii
4	31.a und mehrere 37.b	40.a

[515] Quelle: Eigene Darstellung
[516] Quelle: Eigene Darstellung

Mit den Fragen 7 und 31 wurde im Ansatz bereits auf die Erkenntnisse aus anderen Kategorien zurückgegriffen. Es liegt im Entscheidungsbereich des Datenanalysten weitere Fragen einfließen zu lassen.

In dieser Reifegradkategorie und für die Anwendung der diagnostischen Analyse, liegt der Schwerpunkt auf den Ergebnissen in der Spalte Allgemein.

Aggregation der Bewertung

Für die Zusammenführung der Gesamtbewertungen der einzelnen Reifegradkategorien wird für die Prozessphase Durchführung die Bewertung der Kategorie Datendarstellung vor dem Datenumfang am höchsten gewichtet. Die Darstellung der Daten beeinflusst maßgeblich den Arbeitsaufwand in der Datenaufbereitung. Des Weiteren verliert selbst der umfangreichste Datenbestand seinen Wert, wenn dieser nur aus unstrukturiertem Fließtext besteht.

In der Prozessphase Operationalisierung wird die Kategorie Datenbereitstellung am höchsten gewichtet gefolgt von den Datenformaten. Um die Zusammenhänge umfassend zu betrachten, ist es von Vorteil unterschiedliche Quellen einheitlich auswerten zu können, daher wird das höchste Gewicht auf die Datenbereitstellung gelegt. Es müssen jedoch die Daten in einem auswertbaren Format vorliegen, weshalb den Datenformaten gegenüber der Datenerfassung der Vorzug gegeben wird. Nachdem die diagnostische Analyse nicht als Echtzeitanwendung gedacht ist, wird die Datenerfassung am geringsten gewertet. Für den Datenanalyseprozess in Summe wird die Prozessphase Durchführung über die Operationalisierung gewichtet. Die Ausgestaltung der Daten ist wichtiger als die Art der Aufzeichnung und Bereitstellung. Abgeleitet aus den Gewichtungen und den Reifegradbeschreibungen, ist für die Schwachstellenanalyse das House of Data Quality in Abbildung 40 dargestellt. Die Abschnitte 7.1.1, 7.1.2, 7.1.3 und 7.2.2 beschreiben die Fallbeispiele die schrittweise zur Bewertung der Zusammenhänge im Dach und in der Korrelationsmatrix führten.

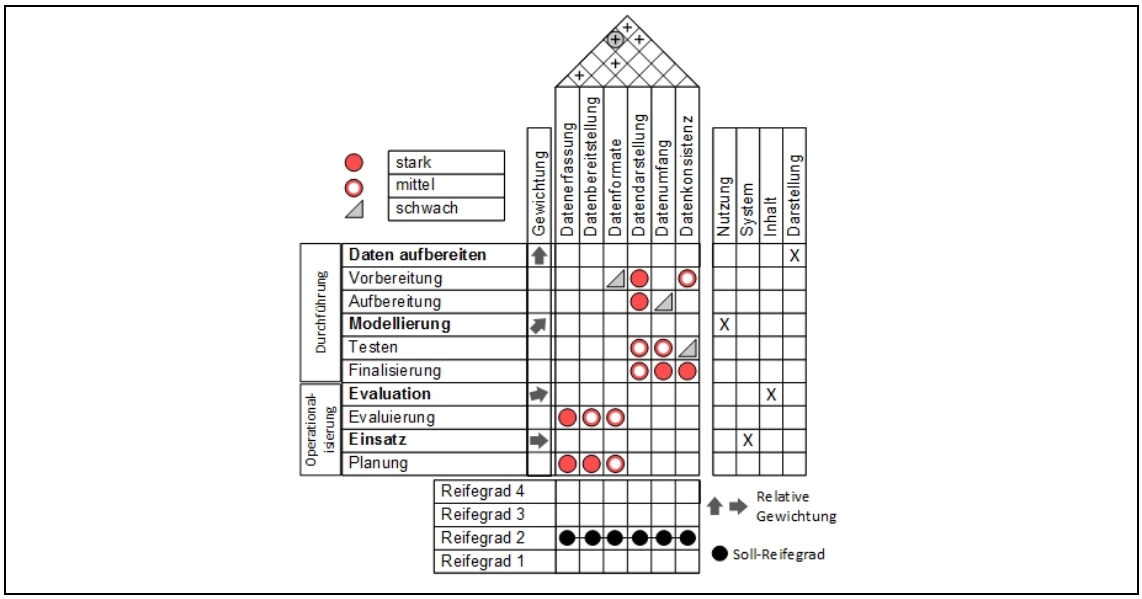

Abbildung 40: House of Data Quality für diagnostische Analysen[517]

[517] Quelle: Eigene Darstellung

6.5 Zusammenfassung Reifegradmodell

Die schnelle Entwicklung von Industrie 4.0 und die damit steigende Menge an Daten motiviert Unternehmen verstärkt Projekte in der Datenanalyse durchzuführen. Es fehlte bisher jedoch an geeigneten Mitteln festzustellen ob eine Organisationseinheit die nötige Prozessreife erreicht hat, um ein Datenanalyseprojekt zu starten. Das hier entwickelte Reifegradmodell schließt diese Lücke.

Durch die Einbeziehung eines generischen Datenanalyseprozesses wird eine allgemeingültige Anwendbarkeit gewährt. Zur Bewertung von Unternehmen mit datenanalytischen Problemstellungen, die über die diagnostische Analyse hinausgehen, wurden die Prozessphasen verallgemeinert.

Die Berücksichtigung von bereits vorhandenen und - wo nötig - angepassten Datenqualitätsdimensionen gewährleistet, dass die etablierte Theorie in der Datenqualitätsmessung, im Modell ihre Anwendung findet. Sie bildet eine universelle Schnittstelle zwischen der generischen Festlegung des Analyseprozesses und der empirischen Erhebung durch die Reifegradkategorien.

Die sechs Reifegradkategorien sind auf eine Weise gewählt und definiert, dass sie den Prozess der analytischen Datenverarbeitung von der Erfassung bis zu den analysespezifischen Qualitätsanforderungen abbilden. Damit werden spezifische und in der Praxis umsetzbare Handlungsanweisungen geben, die auf der Ebene der Prozessphasen oder der Datenqualitätsdimensionen nicht vorhanden sind.

Die Reifegradstufen wurden in deren hierarchischen Ausgestaltung so gewählt, dass sie die steigende Komplexität und den zunehmenden Nutzen für ein Unternehmen der Analysen abbilden. Jeder Reifegrad bildet eine logische Basis für die übergeordneten Reifegrade. Dadurch wird eine kontinuierliche Entwicklung ermöglicht.

Die Bewertung erfolgt mit einem umfangreichen, jedoch zweckgenauen Fragebogen unterstützt durch Analysen basierend auf den Datenqualitätsdimensionen. Durch die Zuordnung von eindeutigen Ausprägungen in den Fragen zu den Reifegradkategorien und Reifegradstufen dienen diese als wichtige Standardisierung für die weitere Bewertung. Damit soll eine Vergleichbarkeit in den Bewertungen unterschiedlicher Fallbeispiele gewährleistet werden.

Die Entwicklung des Reifegradmodells erfolgte mit dem Design Science Research Ansatz. Dieser hat die Ausarbeitung der Artefakte an Fallbeispielen als zentrale Umsetzungsstrategie. Die Fallbeispiele und deren Beitrag und folglich die Weiterentwicklung des Reifegradmodells werden im folgenden Kapitel beschrieben.

7 Anwendung und Entwicklung des Modells anhand von sechs Fallstudien

Die Entwicklung des Reifegradmodells erfolgte anhand von sechs Projekten, wobei die Anwendung im Rahmen der Schwachstellenanalyse an drei Projekten erfolgte, die hier genauer beschrieben werden. Die anderen drei Projekte lieferten Erkenntnisse für die Entwicklung und Validierung einzelner Details. Tabelle 20 gibt eine Übersicht über die Anwendungsfälle in den Projekten. Die hellgrau hinterlegten Zeilen sind jene Projekte, an denen das Modell exemplarisch anhand der Schwachstellenanalyse entwickelt wurde.

Tabelle 20: Übersicht Fallstudien

Unternehmen	Projektinhalt	Beitrag zum Reifegradmodell
Automobilproduzent – Motorenwerk	Entwicklung eines Instandhaltungsleitstandes unterstützt mit Datenanalytik	Initiale Idee für das Reifegradmodell. Grobe Abgrenzung der Reifegradkategorien
Stahlverarbeiter – Industriegüter	Prognose des Wartungszeitpunktes aufgrund der Drehmomentveränderung	Erste Bestätigung des Grobaufbaus und die Verfeinerung zu den sechs Reifegradkategorien
Textilunternehmen	Analyse von Ausfällen zur Ursachenfindung	Anwendung der Assoziationsanalyse zur Schwachstellenanalyse. Erste Bewertung mit dem Reifegradmodell und folglich Verfeinerung der Reifegradkategorien und –stufen
Automobilzulieferer – Fremdmontage	Modell zur Anomalieerkennung	Bewertung jenseits der Schwachstellenanalyse. Weitere Verfeinerung. Handlungsempfehlung und erfolgte Verbesserung des Reifegrades.
Bau und Montage von Hausgeräten	Entwickeln eines Prozesses für die Schwachstellenanalyse	Handlungsempfehlung zur Erhöhung des Reifegrades. Anwendungstest für klassische Schwachstellenanalyse
Holzverarbeiter – Fußbodenhersteller	Finden von Ursachen für Prozessschwachstellen	Durchgängiger Einsatz des Reifegradmodells. Einsatz für die Schwachstellenanalyse. Input für die Finalversion. Festlegung der Reifegradstufen.

7.1 Generischer Input für das Reifegradmodell

Folgend werden jene Fallbeispiele beschrieben, die einen generischen Beitrag zum Reifegradmodell lieferten. Anhand der definierten Metaphasen des CRISP-DM werden die Fallbeispiele erörtert und der Beitrag zum Reifegradmodell dargestellt.

7.1.1 Fallbeispiel Automobilproduzent

Beim Unternehmen handelt sich um einen international agierenden Automobilproduzenten mit einem Werk für Dieselmotoren in Österreich. Am österreichischen Standort wurde ein Projekt zur Instandhaltung in Industrie 4.0 durchgeführt. Die Projektlaufzeit war von April 2014 bis Juni 2017. Im Rahmen des DSR wurde mit diesem Fallbeispiel das vorläufige Design initiiert mit dem in den folgenden Projekten weitergearbeitet wurde.

Vorbereitungsphase

Ziel des Projektes war die Vorhersage von Anlagenausfällen basierend auf dem Produktionsprogramm. Da keine direkten Messungen an der Anlage und den Anlagenkomponenten durchgeführt wurden, gab es den Versuch die Abnützung an Bauteilen oder Werkzeugen über Messungen der Produktqualität zu erkennen. Die Produktqualität wurde durch eine Stichprobenmessung der Fabrikate nach dem Abschluss der Fertigungsschritte des untersuchten zerspanenden Bearbeitungszentrums festgestellt. Über die Produktqualitätsdaten, sollte ein Muster über die Abnutzung der Komponenten in den Instandhaltungsdaten gefunden werden. Das gefundene Muster sollte über das Produktionsprogramm abbildbar sein, um bereits im Vorfeld die erwartete Abnützung simulieren und Ausfälle vorhersagen zu können. Durch die Hinterlegung von Kosten sollte der ideale Austauschzeitpunkt bestimmbar sein, der neben den Ersatzteilen auch Kombinationseffekte berücksichtigen soll, die sich durch eine Bündelung von Instandhaltungsmaßnahmen ergeben soll.

Die Simulation von Vorgängen und der Vorschlag von idealen Handlungsempfehlungen entsprechen der präskriptiven Analyse.[518] Folglich ist der höchste Reifegrad erforderlich, um die Ziele des Projektes zu erfüllen und die datenanalytische Problemstellung allumfänglich zu lösen.

Es standen drei Datenquellen zur Verfügung. Die Instandhaltungsaufzeichnungen, die Qualitätsmessungen an den Produkten und das Produktionsprogramm der Produktionsplanung.[519] Davon unabhängig gab es Aufzeichnungen über die Stromaufnahme beim Verfahren einer instandhaltungsrelevanten Anlagenkomponente. Der Aufzeichnungszeitraum beschränkte sich auf wenige Wochen und diente als Input für eine Machbarkeitsstudie.

[518] Vgl. Matyas, K. et al. (2017), S. 462.
[519] Vgl. Geißler, P. et al. (2016), S. 16 f.

Durchführungsphase

Die Durchführungsphase wurde stark von der Art der Datenerfassung und der Fragmentierung der Datenhaltung beeinflusst. Die Instandhaltungsdaten mussten manuell aufbereitet werden. Die Rückmeldungen erfolgten schichtgenau in Form von unstrukturiertem Text auf Papier. Diese Meldungen wurden dann in das ERP-System übertragen. Mit Domänenexperten wurden die Textrückmeldungen von 2,5 Jahren analysiert und einer neudefinierten Anlagenstruktur zugewiesen. Die Anlagenstruktur gliederte jedes Bearbeitungszentrum in Modul, Baugruppe und Teil. Des Weiteren wurden die Störungsart und eine mögliche Ursache aus der textuellen Rückmeldung extrahiert. Diese umfangreichen Aufbereitungsschritte wurden von Projektpartner durchgeführt. Tabelle 21 gibt einen Überblick über die eingeführte Standardisierung. Die Instandhaltungsdaten lagen somit nominal skaliert vor. Die Zeitstempelqualität ist fraglich, da der Übertrag ins System nicht sofort erfolgte und nur auf die Schicht genau vorliegt.

Tabelle 21: Störungszuordnung Automobilproduzent[520]

Störungsart		Ursachen
Grob	Fein	
Mechanisch	Achse	Ablaufstörung
Elektrisch	Hardware	Defektes Teil
Sonstige	Hydraulik	Einstellung
	Kühlmittel	Mensch
	Mechanisch	Reset
	Pneumatik	Verschleiß
	Software	Verschmutzung
	Nicht zuordenbar	Werkzeugtausch
		Nicht zuordenbar

Die Qualitätsdaten lagen metrisch skaliert vor. Es handelte sich um standardisierte Messungen am Produkt. Pro Schicht wurde ein Produkt von jedem der Bearbeitungszentren entnommen und an mehreren definierten Messpunkten vermessen. Problematisch war hier ebenfalls die Zeitstempelkonsistenz. Der Zeitpunkt im Messprotokoll entsprach nicht jenem der Entnahme aus dem Produktionsprozess, bzw. noch besser jenem der Fertigung, sondern es war der Zeitpunkt der Messung. Da der Zeitunterschied zwischen der Fertigung und der Messung nicht bekannt oder konstant war, konnte kein definierter Offset entfernt werden, um eine eindeutige Zuordnung vornehmen zu können. Leerstellen existierten nicht, jedoch waren Ausreißer zu finden, die auf die manuelle Übertragung des Messprotokolls zurückzuführen waren.

Die Modellierung sollte mittels eines überwachten Lernverfahrens erfolgen. Dazu waren gelabelte Daten nötig. Die Attribute waren Messungen der Qualitätsdaten, die Labels die Instandhaltungsdaten. Die Modellierung konnte nicht erfolgreich durchgeführt werden,

[520] Quelle: Eigene Darstellung

da es zu wenige ausfallsrelevanten Vorfälle gab. So lagen für ein Bearbeitungszentrum in 2,5 Jahren nur 29 Ausfälle vor, die auf einen Verschleiß mit mechanischer Störungsart zurückzuführen waren. Dazu musste eine weitere Filterung in der Anlagenstruktur erfolgen. So gab es zwölf verschiedene betroffene Teile, wodurch je Teil in 2,5 Jahren nur zwischen zwei und drei Verschleißausfälle existieren. Es sind zu wenige Fälle, um ein Prädiktionsmodell zu trainieren.[521] Die Produktionsdaten wurden aus unterschiedlichen Systemen zur Verfügung gestellt und enthielten je nach System Informationen über die Stückzahlen nach Tag, Schicht und Anlage. Aufgrund der Inkonsistenzen in den Daten wurden diese nicht weiter betrachtet.

Die Stromaufnahmedaten lagen nur als Bilddatei vor und mussten mit einem Bilderkennungsverfahren in Messwerte konvertiert werden. Mit einer Merkmalsextraktion wurden spezifische Charakteristika der Messkurve gewonnen. Weitere Analysen wurde nicht durchgeführt.

Operationalisierungsphase

Die Operationalisierungsphase zeigte die Schwächen der Datenaufnahme und Speicherung. Die Aufzeichnung erfolgte in vielen Fällen manuell auf Papier, mit einem starken Zeitverzug zur Dateneingabe in das System, wobei der gespeicherte Zeitstempel nicht immer mit den relevanten Prozesszeiten übereinstimmte. Durch die Art der Datenspeicherung und Datenerfassung, ist das System nicht für eine automatische Industrie 4.0 Anwendung geeignet und ein datenanalytischer Analyseprozess in Sinne der präskriptiven Analyse ist nicht umsetzbar.

Des Weiteren zeigte sich der immense Aufbereitungsaufwand, der durch die Art der Datenerfassung und Datenspeicherung entstand. Die Instandhaltungsdaten wurden gleichzeitig mit mehreren Personen manuell in ein strukturiertes Format gebracht. Dafür ist Arbeitszeit von Personenwochen angefallen. Das Gleiche galt für die Daten der Qualitätsmessung. Die Datenquelle konnte nicht automatisch ausgewertet werden. Daher mussten die Messungen manuell in Excel übertragen werden. Auch hierzu war Arbeitszeit von Personenwochen nötig. Mit diesen Daten wurde die weitere Datenaufbereitung für die Modellierung durchgeführt, wiederum mit mehreren Personenwochen Aufwand. In Summe ist für die Datenaufbereitung ein Arbeitsaufwand von Personenmonaten angefallen.

Viele der Instandhaltungsdaten konnten nicht eindeutig einer Störung oder einer Ursache zugeordnet werden. Darüber hinaus ist die Glaubwürdigkeit der Datenquelle und Erfassung gering. Durch tlw. proprietäre Formate und Systemrestriktionen waren Daten im Laufe des Projektes nicht mehr zugänglich oder bearbeitbar.

Beitrag für das Reifegradmodell

Dieses Fallbeispiel initiierte die Idee für eine Methode zur Vorabfeststellung der Machbarkeit und der groben Abschätzung des Aufwandes für datenanalytische Projekte. Es zeigte die Relevanz des Datenmanagements mit der Aufzeichnung und der Speicherung der Daten und die Wichtigkeit der Gestaltung der Daten im Skalenniveau und im Umfang für weitere Analysen. Abbildung 41 ist stellt den initialen Zusammenhang

[521] Vgl. Bernerstätter, R. et al. (2016), S. 27 f.

zwischen den Reifegradkategorien und den Prozessphasen her. Die Zusammenhänge zur Datenkonsistenz sind noch nicht eindeutig erkennbar. Das Bild wird sich in den weiteren Fallbeispielen schärfen.

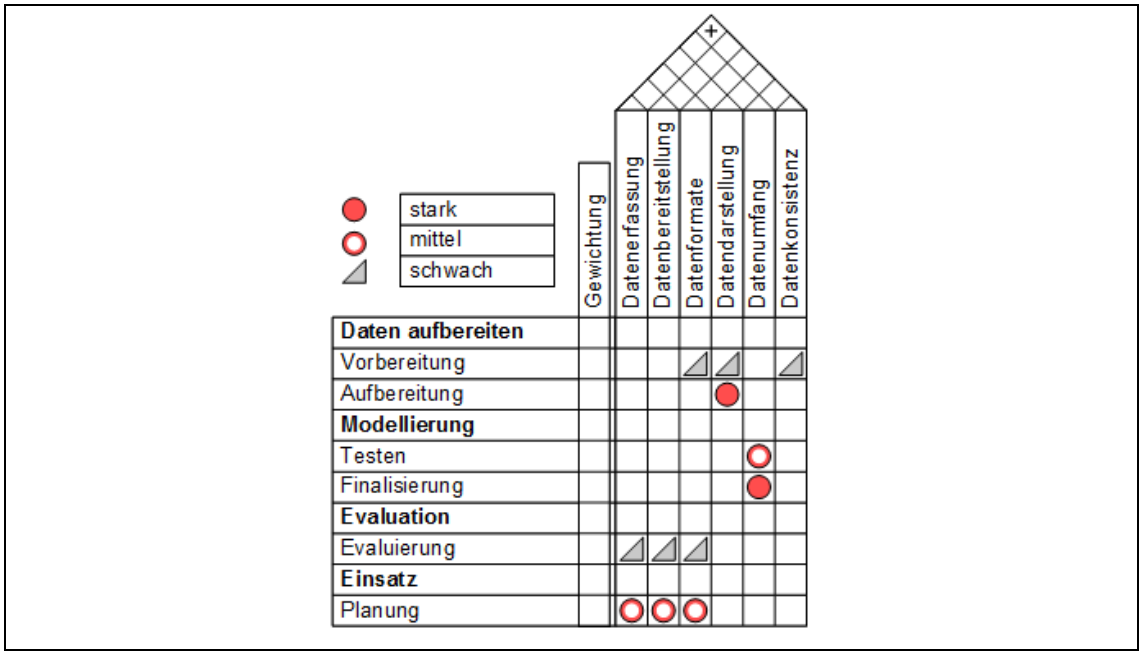

Abbildung 41: Vorläufiges HoDG für Automobilproduzent[522]

Des Weiteren zeigte sich, dass prädiktive Analysen oder präskriptive Analysen einen wesentlich höheren Reifegrad benötigen als normale visuelle Auswertungen. Durch die lange Projektlaufzeit wurden die Reifegradkategorien am Ende des Projekts fixiert. Tabelle 30 zeigt den Beitrag, den das Projekt zu den Kategorien lieferte, die dem Datenmanagement zuordenbar sind. Im Wesentlichen kann aus Tabelle 30 ein geringer Reifegrad zwischen eins und zwei abgelesen werden. Tabelle 31[523] zeigt den Beitrag des Projektes zu den Kategorien, die der Datenstruktur und der Durchführungsphase zuzuordnen sind. Die Datendarstellung ist im nicht aufbereiteten Zustand Reifegrad eins zuzuordnen. Die anderen beiden Kategorien sind zwischen Reifegrad eins und Reifegrad zwei.

7.1.2 Fallbeispiel Stahlverarbeiter

Das Unternehmen ist ein auf Stahlverarbeitung basierender Technologiekonzern, der international tätig ist. Sein Hauptsitz befindet sich in Österreich. Die Produktpalette umfasst Industriegüter von Stahlblechen über Draht bis zu Schienen. Im Rahmen eines umfangreichen Instandhaltungsprojektes wurde ein Anwendungsfall zur prädiktiven Instandhaltung umgesetzt. Das Projekt wurde primär im zweiten Halbjahr 2016 abgewickelt.

Im Rahmen des DSR wurde mit diesem Fallbeispiel das vorläufige Design aus dem Projekt mit dem Automobilhersteller verfeinert und die Reifegradkategorien fixiert. Außerdem gab es Input für die genauere Ausgestaltung der Reifegradstufen.

[522] Quelle: Eigene Darstellung
[523] Tabelle 30 und Tabelle 31 sind in Anhang 1B zu finden.

Vorbereitungsphase

Das übergeordnete Ziel des Fallbeispiels war die Anpassung der präventiven Instandhaltungsstrategie, in diesem Fall der Reinigung eines Walzgerüstes zu fixen Zeitpunkten (FTM). Die Reinigung sollte durchgeführt werden, wenn es prozessbedingt nötig ist. Die Frage, die sich stellte, war das „Wann“ des verschmutzungsbedingten Ausfalls.

Datenanalytisch handelt es sich um eine klassische Prädiktion mit einem überwachten Lernverfahren. Es sind unabhängige beschreibende Attribute und eine abhängige Variable nötig, die von den unabhängigen Attributen beschrieben wird. Im vorliegenden Fall wird die Veränderung des Drehmoments beim Öffnen und Schließen des Walzgerüstes gemessen. Darüber hinaus liegen Aufzeichnungen über die Instandhaltungstätigkeiten vor. Das datenanalytische Modell muss die beiden Datenbestände in Verbindung setzten, um aufgrund der Sensormessungen die Reinigungszeitpunkte vorherzusagen.

Die Sensormessungen sind in Einzeldateien organisiert. Jede Datei enthält die Daten eines vollständigen Walzgerüstwechsels. Dabei werden im Takt von 5ms 46 Signale erfasst. Die Instandhaltungstätigkeiten werden auf der Ebene des technischen Platzes (Walzgerüst) mit einem sekundengenauen Zeitstempel erfasst.

Durchführungsphase

Die Sensoraufzeichnungen mussten im ersten Schritt aus dem Leitstandsystem exportiert werden. Das Datenformat war proprietär und konnte nur mit einem speziellen Programm in eine lesbare Textdatei exportiert werden. In weiterer Folge wurden die Einzeldateien zu einer Gesamtdatei zusammengeführt, um einen Gesamtdatenbestand der Gerüstwechsel eines halben Jahrs zu erhalten. Abhängig vom Produktionsprogramm gab es jeden Tag im Schnitt fünf solche Wechsel. Folglich wurden ca. 900 Dateien manuell exportiert und konvertiert. Die Zusammenführung erfolgt mit einem dafür erstellten Programm in MATLAB.

Nach diesem Schritt konnte in der Datenaufbereitung mit den eigentlichen Aufbereitungsschritten begonnen werden. Dazu wurden mit unterschiedlichen Verfahren Merkmale aus den Messkurven extrahiert, mithilfe derer eine Prädiktion in der Modellierungsphase durchgeführt wurde. Die Merkmale wurden mit den Instandhaltungsaufträgen abgeglichen. Die Instandhaltungsaufträge wurden nur drei Monate gespeichert, weshalb nur ca. die Hälfte der Sensormessungen verwendet wurden. In den drei Monaten wurden 38 projektzielrelevante IH-Tätigkeiten durchgeführt. In Summe benötigte die Datenaufbereitung zwei Personenwochen.

Operationalisierungsphase

Die Operationalisierungsphase zeigte die Schwäche in der Bereitstellung der Daten in Form der Datenquellen und Datenformate. Die Messdaten wurden in der Datenbank des Steuerstandes abgelegt und konnten nur dort geöffnet und in ein bekanntes Dateiformat exportiert werden. Nachdem eine Zusammenführung der Daten durch einen ETL-Prozess nicht möglich war, musste ein eigenes Programm dafür geschrieben werden.

Die Ergebnisse wurden mit den Domänenexperten als nachvollziehbar und repräsentativ eingestuft. Dieses Ergebnis war zu erwarten, da die Glaubwürdigkeit der Datenquellen

als hoch einzustufen war. Durch die Schnittstellenproblematik konnte das Modell jedoch nicht implementiert werden.

Beitrag für das Reifegradmodell

Im Rahmen dieses Projektes wurde die erste Veröffentlichung verfasst, die das Reifegradmodell als Teil enthielt. Es zeigte sich die Notwendigkeit, die Datenquellen, -bereitstellung und -formate als Kategorie zu trennen. Außerdem zeigte sich der Unterschied zwischen der Durchführungs- und der Operationalisierungsphase. Die Modellerstellung war möglich, nachdem die Daten aus dem proprietären System extrahiert wurden. Der Einsatz des Modells in weiterer Folge war nicht möglich, da eine Einbindung durch die Schnittstellenproblematik unmöglich war. Die fehlende Reife schlug sich in der Datenaufbereitung nieder. Abbildung 42 zeigt das verfeinerte HoDG durch das Fallbeispiel des Stahlverarbeiter. Der Zusammenhang zwischen Datenerfassung und Datenumfang bestätigte sich. Darüber hinaus ist dieser Fall ein gut nachvollziehbares Beispiel für den positiven Effekt einer automatischen und digitalen Datenerfassung auf die Konsistenz der Daten. Eine bessere Datenkonsistenz ermöglicht es wiederum bessere Modelle zu erstellen

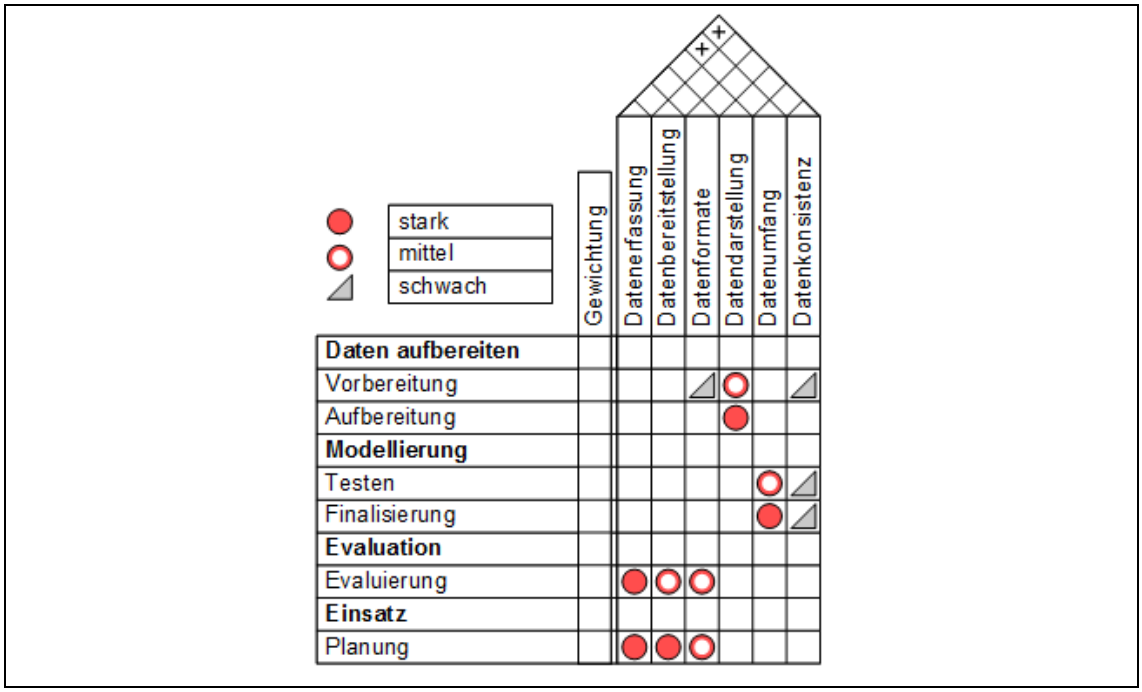

Abbildung 42: Verfeinertes HoDG des Stahlverarbeiters[524]

Tabelle 32 zeigt den Beitrag und die Einordnung des Fallbeispiels zum Reifegradmodell. Ein wesentlicher Punkt ist das proprietäre System und Format. Hier konnte eindeutig die Zweiteilung zwischen den Kategorien für die Operationalisierung und jener für die Durchführung vorgenommen werden. Die Daten selbst können mit dem nötigen Aufwand in die Form gebracht werden, die für die Modellbildung benötigt wird. Eine Implementierung ist jedoch nicht möglich und erschwert zusätzlich den Aufwand in der Aufbereitung.

[524] Quelle: Eigene Darstellung

Tabelle 33 zeigt den Beitrag des Beispiels zur Datenstruktur. Durch die regelmäßigen Wartungen ist die Stichprobe von Ereignissen groß genug, um die Veränderungen in den Sensoraufzeichnungen durch einen überwachten Lernalgorithmus abbilden zu können.[525]

Der CRISP-DM wurde bis zur Evaluation durchlaufen. Zu dem Zeitpunkt des Projekts existierte noch kein strukturierter Fragebogen oder Analysen, um die Daten einzuschätzen. Es wurde mit dem Domänenexperten jedoch ausführlich gesprochen und die Daten vorab auf Verwendbarkeit visuell analysiert. Es wurde daher eine post-hoc eine Einordnung basierend auf dem finalen Reifegradmodell vorgenommen. Tabelle 22 zeigt den Reifegrad für das Fallbeispiel.

Tabelle 22: Reifegradeinordnung Stahlverarbeiter[526]

Reifegrad	Erfassung	Bereitstellung	Formate	Codierung	Umfang	Konsistenz	Gesamt
4						X	
3					X		
2	X	X	X	X			X
1							

Der Gesamtreifegrad für die Durchführungsphase befindet sich bei Reifegradstufe drei, während sich der Reifegrad für die Operationalisierungsphase bei zwei befindet. Die Erstellung eines prädiktiven Modells wäre unter erschwerten Bedingungen möglich. Die Instandhaltungsaufzeichnungen müssten erst aufwendig analysiert werden. Die Implementierung ist nicht möglich, da die Daten dem Modell nicht automatisch zugeführt werden können.

7.1.3 Fallbeispiel Automobilzulieferer

Es handelt sich um einen international agierenden Automobilzulieferer mit 340 Produktionsbetrieben und 89 Kompetenzzentren in 27 Ländern. Der Standort in Österreich bietet Lösungen für Nischenprodukte von Massenfertigung für den internationalen Export von Fahrzeugkomponenten bis hin zur Assemblierung ganzer Fahrzeuge. Das Projekt wurde über ein Jahr (August 2017 bis September 2018) abgewickelt.

Im Rahmen des DSR wurde in diesem Fallbeispiel das Reifegradmodell eingesetzt, um aufgrund der Bewertung Handlungsempfehlungen abzugeben. Die Reifegradkategorien waren zu dieser Zeit fixiert und die Reifegradstufen großteils ausformuliert. Nach Abschluss des Projekts wurden die Kategorien dem CRISP-DM zugeordnet.

Vorbereitungsphase

In Rahmen der Vorbereitungsphase wurde mit einer ersten Version des Fragebogens gearbeitet, der der Datenerhebung zur Reifegradfestlegung dient. Ziel des Projektes war die Vorhersage der Ausfälle von Kompressoren einer Kältemaschine. Die Motivation dahinter war, dass die Anlage nur präventiv mit einem engmaschigen Intervall gewartet

[525] Tabelle 32 Tabelle 33 zu finden in 1C.
[526] Quelle: Eigene Darstellung

wurde. Das führte zu hohen Kosten. Zusätzlich war die Anlage kritisch, da sie zu Kühlung der Prozessleittechnik diente. Des Weiteren sollte eine umfangreiche und kostspielige Überholung stattfinden. Wenn die Überholung durch die Überwachung aufgeschoben werden könnte, wäre das Projekt als Erfolg zu bezeichnen. Die letzte Motivation ist strategischer Natur. Das Unternehmen will verstärkt auf prädiktive Instandhaltung und Datenanalytik setzen. Mit dem Projekt sollten Erfahrungen gesammelt und mögliche Verbesserungspotenziale aufgezeigt werden.

Datenmanagement:

Als Datengrundlage standen für das Fallbeispiel Sensormessungen einer Vibrationsmessung sowie einer Strommessung zur Verfügung. Die Vibrationssensoren wurden über ein Condition-Monitoring Modul aufgezeichnet, welches eine erste Vorverarbeitung in Echtzeit durchführte.

Die Vibrationsmessung wurden dabei in charakteristische Frequenzbänder geteilt, um typische Fehler eines Elektromotors zu identifizieren.[527]

Die Strommessung erfolgte 3-phasig ohne eine weitere Vorverarbeitung der Daten. Die Kommunikation zwischen den Modulen lief über das OPC-UA Protokoll. Ein gesondertes Netzwerk übertrug Steuerungsdaten und Betriebsdaten. Diese enthielten Informationen über den Status der Anlage, wie z. B. Störmeldungen. Diese Störmeldungen können mit den SAP-Aufträgen der Instandhaltung abgeglichen werden.

Die Sensormessungen waren auf einem SQL-Server gespeichert, die Instandhaltungsdaten in SAP. Die verwendeten Programme und Formate waren Standards und verfügten über Schnittstellen, bzw. sind von unterschiedlichen Programmen importierbar.

Datenstruktur:

Die Sensordaten waren metrisch und die Steuerungsdaten nominal-binär skaliert. Die Sensordaten wurden aggregiert auf eine Sekunde und in Frequenzbänder aufgeteilt gespeichert. Die Instandhaltungsdaten waren unstrukturierter Text. Die Erfassung erfolgte nur auf Ebene der Kältemaschine. Um die Tätigkeiten auf Ebene eines technischen Platzes zuzuordnen, mussten die Tätigkeiten über den Zeitstempel mit den automatisch erfassten Daten abgeglichen werden. Metadaten waren nicht automatisch hinterlegt, es waren jedoch die gültigen Messbereiche bekannt, wodurch Ausreißer und Fehlmessungen identifiziert werden konnten. Das Vorhandensein dieser Information war ein Hinweis einer bewussten Datenmodellierung und eines hohen Reifegrades.

Der Datenbestand der Sensoraufzeichnungen betrug bei Projektstart nur zwei Monate. Relevante Ausfälle sind in der Zeit nicht vorgefallen, da die Anlage aufgrund der engmaschigen Instandhaltung stabil läuft. Aufzeichnungen über die Wartungstätigkeiten wurden lückenhaft geführt.

Die Datenkonsistenz war gegeben, da die Datenaufzeichnung in Echtzeit erfolgte und eine direkte Beziehung zwischen der aufgezeichneten Information und dem Zeitstempel

[527] Vgl. Khazraei, K. (2011), S. 237.

bestand. Die Konsistenz zwischen den Instandhaltungs- und den Störddaten war zu bemängeln, da diese nur indirekt über einen logischen Schluss herzustellen waren.

Voranalyse der Daten:

Die vorhandenen Sensoraufzeichnungen wurden visualisiert, wobei sich zeigte, dass die Aufzeichnung nicht regelmäßig erfolgte. Der Aufzeichnungstrigger wurde gesetzt, wenn es in einem Zeitfenster von 100ms eine Wertveränderung zum Letztwert von 5% erfolgte. Des Weiteren zeigte die visuelle Analyse Spitzen bei der Messung durch den Einschaltvorgang und eine negative Frequenzmessung was auf einen Messfehler hindeutete.

Abgeleitete Schlüsse:

Der geringe Umfang relevanter Instandhaltungstätigkeiten führte dazu, dass ein überwachtes Lernverfahren nicht angewandt werden konnte. Folglich wurde das Projektziel zu einer unüberwachten Anomalieerkennung geändert. Die Prognose wird als mittelfristiges Ziel gesehen. Das Projektvorgehen sollte so aufgesetzt werden, dass die erkannten Anomalien von den Domänenexperten klassifiziert werden. Damit steigt der Umfang der Datensätze, mit denen ein überwachtes Modell gelernt werden kann.

Um den Abgleich der Daten anhand eines einheitlichen Zeitstempels durchführen zu können, wurde als Verbesserung empfohlen, die Daten regelmäßig oder wenigstens immer gleichzeitig zu erfassen. Diese Empfehlung wurde umgesetzt und die Aufzeichnungsfrequenz wurde durchgehend auf ein Intervall von einer Sekunde vereinheitlicht.

Durchführungsphase

Da die Umstellung der Aufzeichnungsfrequenz aus organisatorischen und technischen Gründen länger dauerte, wurde beschlossen zusätzlich mit den bereits aufgezeichneten Daten zu arbeiten, um diese Datenbasis, die am Ende vier Monate groß war, nicht zu verlieren. Dazu wurde ein Programm entwickelt, welches die Daten auf das Intervall von einer Sekunde hochsampelt. Die Erstellung des Programms mit den damit verbundenen Tests der am besten geeigneten Samplingmethode, benötigte ca. 60 Stunden. Dazu kommen noch ca. 20 Stunden für die technischen Anpassungen im Datenmanagement durch die erhöhte Datenmenge. Nach Umstellung der Datenerfassung auf die regelmäßige Abtastung, war das Programm nicht mehr nötig. Wäre der Reifegrad sofort auf dieser Stufe gewesen wäre dieser Aufbereitungsaufwand vermeidbar gewesen.

In den weiteren Bearbeitungsschritten wurde festgestellt, dass ein Statusflag, welches den Betrieb eines Kompressors anzeigt, nicht vollständig synchron mit dem Start und dem Ende des Betriebs war. Zur Verringerung der Datenmenge wurden im Modell nur Daten verwendet, die mit dem Statusflag ‚1' gekennzeichnet waren. Es mussten die Daten daher neu aufbereitet werden, um dem Offset in der Zeitstempelkonsistenz der Daten Rechnung zu tragen. Es war damit ein Aufwand von ca. 20 Stunden verbunden. In der Voranalyse im Datenverständnis wurde dieser Umstand übersehen, da die Daten durch die unterschiedliche Aufzeichnungsfrequenz nicht eindeutig übereinandergelegt werden konnten. Auch diese 20 Stunden Arbeitsaufwand hätten vermieden werden

können, wenn der Reifegrad höher gewesen, bzw. der Umstand des Offsets vorab bekannt gewesen wäre.

Die restlichen Schritte in der Durchführungsphase beinhalteten die Anomalieerkennung und die Erstellung eines Klassifikationsmodells, um die zukünftigen Anomalien zuordnen zu können. Um eine Vorwarnzeit im Sinne einer prädiktiven Instandhaltung geben zu können, wurden vier Anomalieklassen definiert, die den zeitlichen Aspekt einer Anomalieentstehung abbilden sollten.

Operationalisierungsphase

Nachdem das Modell fertig gestellt wurde, wurde das Ergebnis mit den Fachexperten diskutiert. Die Anomalieklassen umfassten die Kategorien:

- OK: Für Messungen die nicht als Anomalie zu sehen sind und den Normalbetrieb darstellten.
- Anbahnung: Für Messungen die auf eine mögliche Entstehung einer Anomalie hindeuten.
- Warnung: Für Messungen die definitiv auf eine Anomalie hindeuten. Der Betrieb der Anlage ist noch möglich, wenn auch der Ausfall in absehbarer Zeit wahrscheinlicher wird.
- Alarm: Es handelt sich um eine Anomalie, die sofortiges Handeln benötigt.

Die Grenzen der Klassen wurden in der Evaluation kritisch diskutiert und mit den Mess- und Instandhaltungsdaten abgeglichen. Da die Aufzeichnungsart und die Datenquellen glaubwürdig waren bildeten sie eine gute Diskussionsgrundlage. Das ERP System erlaubte einen einfachen Zugang für die Diskussion. Das Ergebnis der Diskussion war eine leichte Verschiebung der Klassengrenzen, um das System nicht zu empfindlich zu gestalten. Das Modell musste daher neu trainiert werden, was jedoch hauptsächlich mit Rechenzeit verbunden war und auch mit dem besten Reifegradmodell nicht vermeidbar wäre.

Die Implementierung erfolgte ohne größere technische Schwierigkeiten, da die Systeme und Daten leicht zugänglich waren. Die Art der Datenerfassung und –bereitstellung ermöglichte die Verwendung der Daten in Echtzeit.

Beitrag für das Reifegradmodell

Im Rahmen des Projektes wurden in der Vorbereitungsphase strukturiert das Ziel und die datenanalytische Problemstellung erarbeitet. Aufbauend darauf war bekannt, welche Anforderungen an die Daten notwendig sind. Es zeigte sich auch, dass die Unterteilung des CRISP-DM in die drei Metaphasen sinnvoll ist. Die erste Phase bereitet den datenanalytischen Prozess vor, während die zweite Phase die eigentliche Arbeit eines Data Scientist umfasst. Die dritte Phase kann, muss jedoch nicht umgesetzt werden.

Es zeigt die Notwendigkeit explizit auf die Konsistenz der Daten in der zeitlichen Perspektive zu achten. In den Fragebogen und den Voranalysen im Rahmen des Datenverständnisses fanden diese Erkenntnisse Eingang.

Des Weiteren bestätigte sich der vorab angenommene Zusammenhang zwischen den Kategorien Datenerfassung, -umfang und –konsistenz wobei dieser zwischen Erfassung und Umfang sehr stark ist. Abbildung 43 zeigt darüber hinaus den Einfluss der

Datenkonsistenz auf die Vorbereitung der Datenaufbereitung und die Modellierungsphase.

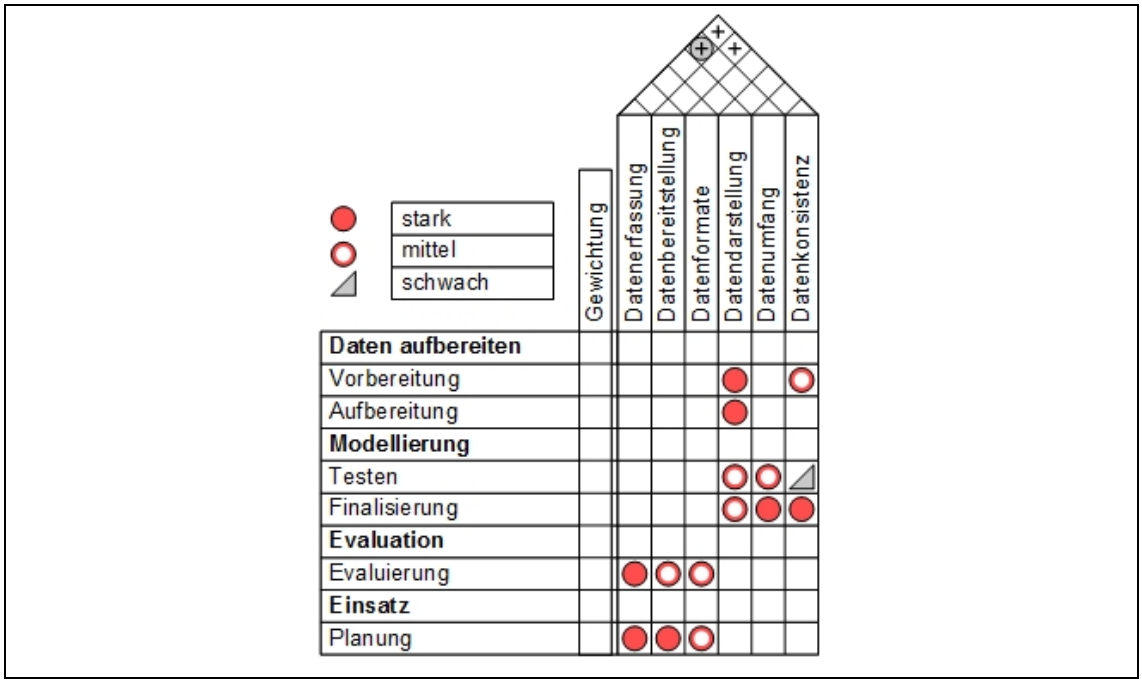

Abbildung 43: Verfeinertes HoDG durch Automobilzulieferer[528]

Durch die Handlungsempfehlung die Datenerfassung regelmäßig (alle Daten im gleichen Takt) aufzuzeichnen wurde der Reifegrad verbessert. Es zeigte sich die theoretische Aufwandsreduzierung bei der Datenaufbereitung und der positive Effekt eines höheren Reifegrades. In Tabelle 23 wird die Reifegradeinordnung vorgenommen. Die Datenerfassung wird mit Reifegrad zwei bewertet, da die Daten nicht regelmäßig erfasst (Sensor- und Prozessdaten) und die Instandhaltungsdaten manuell erfasst werden. Die Datenbereitstellung wurde mit Stufe drei beurteilt, da die Daten in keinem einheitlichen Data-Warehouse abgelegt werden, sondern in unterschiedlichen Systemen. Die Datenformate entsprechen dem dritten und nicht dem vierten Reifegrad, da sie nicht in Big Data Formaten abgelegt sind oder in Formaten mit Metadaten. Die Datencodierung wurde nur mit Stufe 2 bewertet, da die Instandhaltungsdaten als unstrukturierter Text abgebildet sind und die Aufzeichnung auf der obersten Hierarchiestufe zurückgemeldet werden.

Tabelle 23: Reifegradeinordnung Automobilzulieferer[529]

Reifegrad	Erfassung	Bereitstellung	Formate	Codierung	Umfang	Konsistenz	Gesamt
4							
3		X	X			X	
2	X			X	X		X
1							

[528] Quelle: Eigene Darstellung
[529] Quelle: Eigene Darstellung

Der Abgleich auf Ebene der einzelnen Kompressoren ist zwar möglich, unter Umständen jedoch nicht eindeutig. Der Umfang der Daten ist ebenfalls mit Reifegrad zwei bewertet, da es kaum Instandhaltungsaufzeichnungen gab und zu Beginn des Projektes nur ungenügend Sensoraufzeichnungen vorhanden waren. Die Konsistenz ist mit Reifegrad drei bewertet, da aufgrund des Offsets Reifegrad vier nicht erreicht werden kann.

7.2 Spezifische Ausarbeitung anhand Beispiele der Schwachstellenanalyse

Folgende Fallbeispiele haben eine Schwachstellenanalyse zum Ziel. Um solche Analysen durchführen zu können ist der zweite Reifegrad über die sechs Kategorien nötig. Anhand der Fallbeispiele wurde das Reifegradmodell in den Extrempositionen eingesetzt. Ein Fallbeispiel hatte zum Ziel eine einfache Analyse, basierend auf deskriptiven Auswertungen zu ermöglichen. Die beiden weiteren Fallbeispiele umfassten eine datengestützte Schwachstellenanalyse, die über den klassischen Controllinggedanken hinausgehen.

7.2.1 Fallbeispiel Hausgerätehersteller

Das Unternehmen ist ein international tätiges Familienunternehmen mit einer großen Branchenbreite. Die Hausgeräteherstellung hat einen Standort in Österreich und produziert dort Kühl- und Gefriergeräte. Im Rahmen eines Instandhaltungsprojektes wurde die Einführung einer standardisierten Schwachstellenanalyse als eine wichtige Maßnahme zur Verbesserung der Instandhaltung betrachtet. Dieser Teil des Projektes wurde im vierten Quartal von 2018 bearbeitet.

Im Rahmen des DSR wurde in diesem Fallbeispiel bewusst eine Projektsituation bearbeitet, in der das Unternehmen keinen hohen Reifegrad benötigte. Dabei sollte festgestellt werden, ob einfache Analyselösungen ohne eine Big Data Unterstützung korrekt bewertet werden können. Das Vorgehen wurde trotzdem am CRISP-DM ausgerichtet.

Vorbereitungsphase

Im Rahmen eines Instandhaltungsassessments wurde die Schwachstellenanalyse mit dem Reifegrad zwei von fünf möglichen bewertet (Abbildung 44). Die Untergliederung der Reifegrade für die Schwachstellenanalyse wurde in Abschnitt 5.2.2 beschrieben. Im vorliegenden Reifegrad es Fallbeispiels wurde die Schwachstellenanalyse durchgeführt, jedoch ohne Standards im Ablauf oder in den Daten. Als Ziel der Analyse wurde eine einfach deskriptive ABC-Analyse definiert. Dieser Reifegrad zwei der Schwachstellenanalyse entspricht dem ersten Reifegrad des entwickelten Modells. Es werden die Daten der Reparaturen erfasst. Dazu zählen der Ort einer Reparatur, der Fehlercodes und die textuelle Erfassung der Ursache.

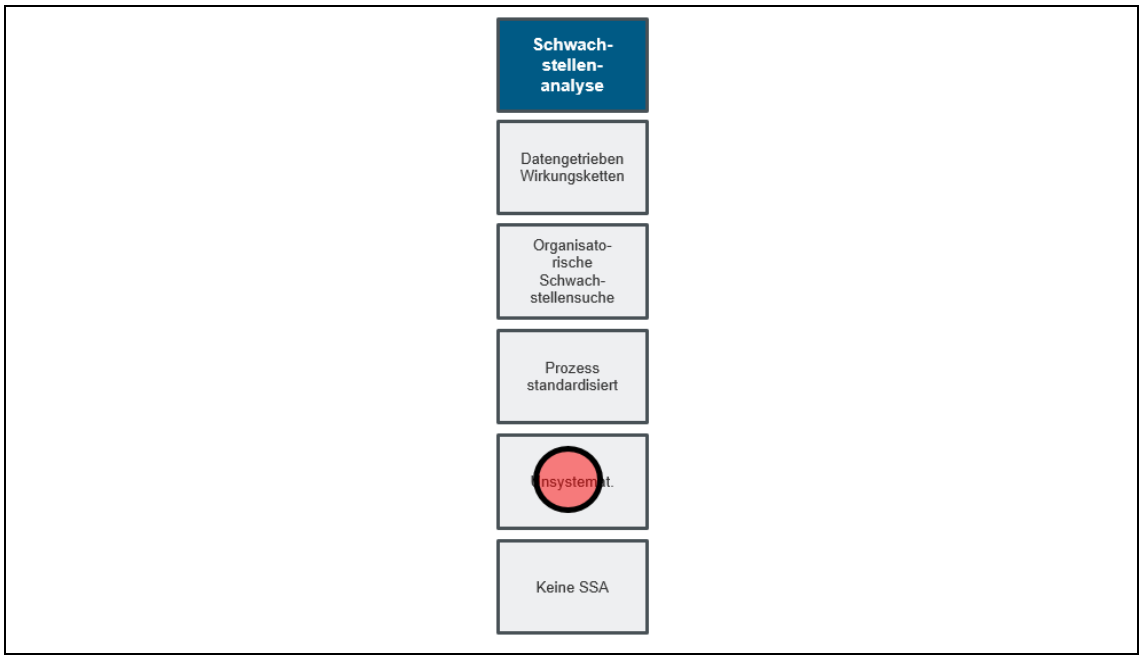

Abbildung 44: Vorabbewertung der Schwachstellenanalyse[530]

Operationalisierung:

Die Datenerfassung erfolgte manuell und nicht digital. Die Unregelmäßigkeit ist durch den Anlagenausfall als Prozesstrigger gegeben. Es gab keine automatisierten Fehlerdetektionssysteme. Organisatorisch wurden die Daten durch den Vorgesetzten überprüft. Er bewertete die Einträge auf Grund seiner Erfahrung, informierte sich bei den Mitarbeitern über Unklarheiten und klärte über die Notwendigkeit einer korrekten Eingabe auf. Die Glaubwürdigkeit wurde mit ‚1' eingestuft, während die Fehlerfreiheit mit ‚2' eingestuft wurde, da es organisatorische Überprüfungsmaßnahmen gibt. Der Reifegrad der Kategorie wurde folglich mit ‚1' bewertet.

Die Daten wurden verspätet in das eigene Instandhaltungssystem übertragen und somit digitalisiert. Das System selbst hatte keine Schnittstellen zu anderen IT-Systemen. Die Analysen mussten darin durchgeführt werden. Es konnte somit als proprietär eingestuft werden. Der Reifegrad war somit mit ‚1' zu bewerten. Die Datenformate waren Standardformate für Tabellenkalkulationsprogramme, in diesem Fall ‚xlsx'. Der Reifegrad wurde mit ‚2' eingestuft.

Es zeigte sich, dass die CRISP-DM Phase der Operationalisierung einen geringen Reifegrad aufwies. Der Gesamtreifegrad wurde mit ‚1' einzustufen. Für die Erfassung großer Datenmengen war das System nicht geeignet und eine automatisierte Analyse nicht möglich.

[530] Quelle: Eigene Darstellung

Durchführung:

Die Datendarstellung zeigte sich in unterschiedlicher Weise. Der Fehlercode war standardisiert und im nominalen Skalenformat. Die Störungsursache ist unstrukturierter Fließtext. Das nominale Attribut Fehlercode umfasste folgende 13 Ausprägungen:

- Ausschäumung
- Bedienfehler
- Betriebsmittelverbesserung
- Elektronik/Hardware
- Elektronik/Software
- Hydraulik
- Materialfehler
- Mechanik
- Pneumatik
- Überlast
- Vakuum
- Verschleiß
- Wartungsmangel

Die genauere Analyse des Fehlercodes ergab, dass diese nicht eindeutig auslegbar waren. Mechanische Fehler können auch an der Hydraulik oder Pneumatik auftreten. Des Weiteren sind einige der Einträge Ursachen für die Fehler, wie zum Beispiel ‚Überlast'. Zur Bewertung der einheitlichen Darstellung wird Formel 6-2 verwendet. Somit ist die einheitliche Darstellung in der Kategorie mit ‚0' zu bewerten.

Zur Bewertung der eindeutigen Auslegbarkeit wurde Formel 6-4 verwendet. Die eindeutige Auslegbarkeit war für beide Attribute mit ‚0' zu bewerten. In Summe wurde der Reifegrad der Kategorie *Datendarstellung und – codierung* mit ‚1' eingeordnet. Die beiden Datenqualitätsdimensionen waren jeweils mit ‚0' bewertet und es gab keine anderen positiven Effekte in dieser Kategorie.

Für die Kategorie Datenumfang wurde zuerst die Vollständigkeit der Eintragungen ermittelt. Es gibt 858 Rückmeldungen in der Reparaturauswertung. Es fehlt ein Eintrag bei den Fehlercodes und 133 bei den Ursachen.[531] Berechnet nach den Formeln 6-6 und 6-7 wird die Vollständigkeit mit ‚0' bewertet. Hauptgrund waren die vielen fehlenden Einträge in den Ursachen.

Die Relevanz von Daten hängt zum Teil von der Aufgabenstellung ab. Für die Schwachstellenanalyse ist es wichtig die Daten an den richtigen Stellen zu erfassen und dass es eine genug große Bandbreite an Einträgen gibt, um sinnvolle Auswertung durchführen zu können. Abbildung 45 zeigt die Verteilung der Rückmeldungen. Diese Abbildung ist unabhängig von der Bewertung der eindeutigen Auslegbarkeit zu sehen. Es zeigte sich, dass die Rückmeldungen eine große Bandbreite an unterschiedlichen Einträgen haben. Eine genauere Analyse nach Formel 6-1 war nicht nötig, da klar ersichtlich ist, dass kein Ungleichgewicht der Attributsausprägungen vorliegt.

[531] Unbekannt und Fragezeichen werden im Sinne der ‚informative missingness' nicht als fehlende Einträge gewertet. Striche, Punkte und sonstige Platzhalter gelten als fehlende Einträge.

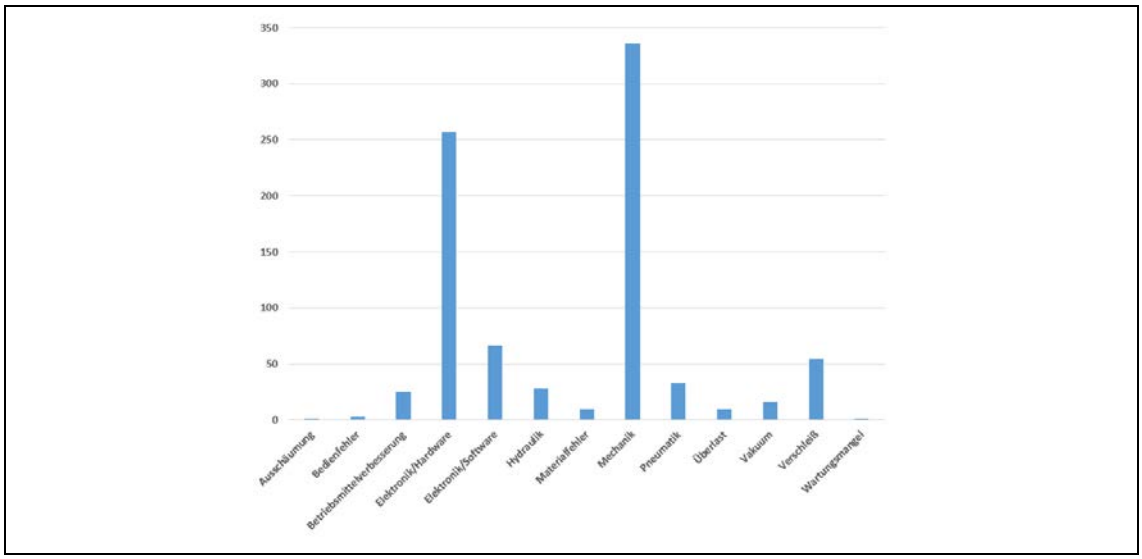

Abbildung 45: Häufigkeit der Rückmeldungen[532]

Die Rückmeldungen konnten auf drei Hierarchiestufen vorgenommen werden. Eine genauere Analyse zeigte, dass von den 858 Rückmeldungen, nur 41% die Ebene mit der größten Granularität rückgemeldet wurden. Auf die zweithöchste Ebene werden 742 Fälle gemeldet, was 86% entspricht. Rückfragen mit den Domänenexperten ergaben, dass diese Verteilung an einer mangelnden Rückmeldedisziplin liegt und am Fehlen der Hierarchiestufen, um eine Rückmeldung zu ermöglichen. Die Relevanz der Daten wird daher mit ‚1' bewertet.

Für die Bewertung standen die Daten von acht Monaten zur Verfügung. Wären nur vollständige Daten verwendet worden also jene 725 Einträge in denen auch Ursachen eingetragen wurden und nur die relevanten, also jene die auf die niedrigste Hierarchiestufe rückgemeldet wurden, würde sich der Umfang noch weiter verringern. Der angemessene Umfang ist somit mit ‚1' bewertet.

Der Reifegrad in der Kategorie Datenumfang wurde zusammenfassend mit ‚1' bewertet. Es waren genügend Daten für eine deskriptive Analyse vorhanden. Für diagnostische Analysen oder darüber hinaus ist der Umfang an Datensätzen und an Attributen zu gering.

Die Kategorie Datenkonsistenz umreißt die Problemstellung der Zeitstempel. Die Aktualität gibt hier an, wie groß der zeitliche Unterschied zwischen der Datenspeicherung und der Datenerfassung, bzw. des Auftretens des zu erfassenden Ereignisses war. Nachdem die Aufzeichnungen erst am Ende des Tages oder der Woche übertragen wurden, ist der Zeitstempel nicht als ausreichend konsistent mit dem Ereignis anzusehen Des Weiteren gibt es im Export der Reparaturauswertung keine Zeitstempel. Der Export erfolgt mit der Angabe eines Zeitfensters. Daher kann keine zeitliche Reihung der Ausfälle vorgenommen werden. Der Reifegrad in der Kategorie ist ‚1'.

[532] Quelle: Eigene Darstellung

Gesamtreifegrad:

Die Reifegradbewertung wird in Tabelle 24 zusammengefasst. Der Gesamtreifegrad der Organisation für datenanalytische Projekte ist mit „1“ zu bewerten. Für die Aufgabenstellung einer Schwachstellenanalyse mit den zur Verfügung gestellten Daten und dem Datenmanagementsystem sind rein deskriptive Analysen möglich.

Tabelle 24: Reifegradbewertung Hausgerätehersteller[533]

	Datenanalyse									III
	Operationalisierung					Durchführung				
Reifegrad	Daten-erfassung	Datenbereit-stellung	Daten-formate			Daten-codierung	Daten-umfang	Daten-konsistenz		
1	X	X		X		X	X	X	X	X
2			X							
3										
4										

Durchführungsphase

Es wurden zwei Handlungsempfehlungen ausgesprochen. Zum einen sollte die Datenerfassung verbessert werden. Dazu müssen die MDE und das Instandhaltungssystem integriert werden. Die MDE erlaubt eine bessere Erfassung in Bezug auf Umfang und Konsistenz.

Zum anderen muss das Codierungssystem überarbeitet werden. Die Fehlercodes müssen Fehler wiedergeben und keine Ursachen. Diese müssen mit der Meldungshierarchie verbunden werden, damit nur mehr für die Hierarchiestufe und den Prozessschritt mögliche Fehler angezeigt werden. Des Weiteren sind für das Unternehmen übliche Fehlercodes und Ursachen zu definieren. Die Verbindung aus MDE Daten und den wahren Ursachen sollte den Reifegrad auf ein Niveau heben, um einfache datengestützte Schwachstellenanalyse durchführen zu können.

Ableitung eines standardisierten Codierungsschemas:

Für die Definition der Fehlercodes (Schadensbilder) wurde im Rahmen eines Workshops über die üblichen Ausfälle diskutiert. Es wurde eine Hierarchie von Fehlercodes mit dazugehörigen Feincodes definiert. Abhängig vom Schadensort, werden nur mehr bestimmte Fehlercodegruppen angezeigt. Es wurden sechs Fehlercodegruppen mit dazugehörigen Feincodes definiert.

[533] Quelle: Eigene Darstellung

Ableitung von Ursachencodes:

Textmining ist ein gängiges Mittel, um aus unstrukturierten Texten vergleichbare Daten zu extrahieren. Es wurde verwendet um passende Standardursachen (Schadensursachen) zu finden, indem die Ursacheneinträge analysiert wurden. Gängige Stoppwörter und Ähnliches wurden entfernt. Aus den Ergebnissen wurden für das Unternehmen relevante Schadensursachen definiert und den Schadensbildern zugeordnet. Werden die Maßnahmen umgesetzt kann in Zukunft ohne eine weitere Datenaufbereitung eine deskriptive Analyse, wie die Pareto-Auswertung, für die Schwachstellenanalyse durchgeführt werden.

Beitrag zum Reifegradmodell

Im Rahmen des Projektes wurde mit einfachen Auswertungen die Datenqualität mit Input der Datenqualitätsdimensionen in der Durchführungsphase bewertet. Daraus abgeleitete Handlungsempfehlungen wurden erarbeitet und befinden sich in der Umsetzung. Die Analyse und die Ausarbeitung umfasste etwa fünf Personentage. Die Implementierung des Codierungssystems soll laut Projektpartner drei Personentage in Anspruch nehmen.

Die Bewertung der Reife in den Reifegradkategorien der Operationalisierungsphase zeigt, dass es mit einem geringen Reifegrad nicht möglich ist, die Analysen zu automatisieren, da die Datenerfassung und die -bereitstellung für die automatisierte Durchführung der Analyse ungeeignet sind. Der Aufwand für die Reifegradverbesserung wird ungleich höher sein. Ein gestartetes Projekt zur MDE Einführung umfasst einen Aufwand mehrerer Personenwochen und dazugehörige Investitionskosten. In diesem Zuge soll das Codierungssytem eingeführt werden. MDE Systeme und Codierungssyteme gehen Hand in Hand, da deren Rückmeldungen standardisiert erfolgen. Das MDE bildet die Basis für höhere Reifegrade und komplexere Analysen der Reifegrade drei und vier.

7.2.2 Fallbeispiel Textilunternehmen

Die Daten dieses Fallbeispiels stammen von einem Produktionsstandort eines Textilherstellers von Baumwollstrickwaren in Fernost. Der Fokus liegt auf Produkte für Neugeborene bis Kleinkinder. Beliefert werden Kunden renommierter Marken in der EU und den USA.

Die Ergebnisse der eingesetzten Assoziationsanalyse wurden in einem datengestützten Ishikawa Diagramm visualisiert. Zusätzlich gab das Fallbeispiel Input zur Abgrenzung der Reifegrade, zu den Analysen in den Reifegradkategorien der Operationalisierungsphase und zur Konkretisierung des Fragenkatalogs.

Vorbereitungsphase

Es sollte festgestellt werden, ob in den großen Datenmengen, die über die Jahre aufgezeichnet wurden Zusammenhänge gefunden werden können, die auf die Entstehung gewisser Fehler hindeuten. Deren Eintreten sollte verhindert bzw. rechtzeitig darauf reagiert werden. Um Regeln zu finden, die die Ausfälle erklären fiel die Entscheidung auf die Assoziationsanalyse.

Zur Verfügung standen Störmeldungen und Maschinendaten. Des Weiteren wurden die Daten über die Aufträge, die Produkte und den Rezepturen zur Verfügung gestellt. Die Rezepturen enthielten die Einstellungsvorschriften und das Maschinensetup. Die Aufzeichnungsdauer umfasste drei Jahre und die Daten wurden von sechs unterschiedlichen Maschinen geliefert.

Operationalisierung:

Alle Daten, die zur Verfügung standen, wurden automatisch aufgezeichnet. Der Maschinenstatus wurde regelmäßig im Abstand von 30 Sekunden erfasst. Störungen wurden mit allen Attributen, die aufgezeichnet werden konnten bei Auftreten erfasst. Aufgrund dieser Voraussetzungen wäre die Reifegradkategorie Datenerfassung mit Reifegrad „4“ zu bewerten. Die Glaubwürdigkeit wird mit „3“ eingestuft, da keine genauen Informationen über das Erfassungsumfeld vorlagen. Die Fehlerfreiheit wurde mit „1“ bewertet, da es keine automatisch logischen oder organisatorischen Verfahren gibt, die die Daten kontrollieren. Der Reifegrad der Kategorie wird daher auf „3“ korrigiert.

Die Daten werden in Echtzeit in Datenbanken auf SQL-Basis übertragen. Die typischen Schritte und Abfragen des ETL Prozesses sind möglich. Die Systeme sind nicht vernetzt, ein Abgleich kann nur mit Hilfstabellen und Abfragen und großem manuellen Aufwand und großem Datenverständnis erfolgen. Der Reifegrad der Kategorie Datenspeicherung wird daher mit „2“ bewertet.

Die Exporte aus den Datenbanken erfolgen im standardisierten csv-Format. Des Weiteren ist es möglich eine direkte Verbindung zum SQL-Server herzustellen. Der Reifegrad ist in der Kategorie Datenformate mit „4“ zu bewerten. Es handelt sich um ein durchgängig einheitliches Datenformat, welches auch große Datenmengen fassen kann.

Die gesamte Operationalisierungsphase wird mit Reifegrad „3“ eingestuft. Für eine diagnostische Analyse ist dieser Reifegrad ausreichend. Durch den geringeren Reifegrad in der Kategorie Datenspeicherung, war in der Datenaufbereitung mit erhöhtem Arbeitsaufwand zu rechnen.

Durchführung:

Die Datencodierung ist vollständig standardisiert. Vorgänge, Prozessschritte, Störmeldungen oder Maschinenstatus sind mit Standardcodes im System hinterlegt und werden auch so erfasst. Messungen und Vorgaben zu Prozessparametern und Anlagenparametern, wie Wärme, Vorschubgeschwindigkeit oder eine Gewebespannung sind metrisch skalierte Daten. Störungs- und Vorgangsdaten liegen jedoch nicht atomar vor. Der Ort, die Art und die Störung selbst sind in einen Code zusammengefasst. Eine Auftrennung ist möglich, da die Teile des Codes standardisiert sind, es ist jedoch mit manuellem Aufwand verbunden.

Für jede der sechs Anlagen sind die Störungen in sechs Standardgruppen - abhängig von der unmittelbaren Auswirkung auf den Produktionsprozess - kategorisiert. Störungstyp sechs umfasst generelle Warnungen, die noch zu keinem Stillstand führen. Ab Störungstyp vier wird die Produktion gestoppt, wobei es sich um unterschiedliche Eskalationsstufen handelt. Störungstyp drei fährt die Anlage langsam innerhalb von 10 Sekunden herunter, Störungstyp zwei innerhalb von 3 Sekunden und Störungstyp eins

umfasst alle Störungen die einen Notstopp auslösen und zu Folgeschäden führen können. Innerhalb der Typen sind die Störungen in Subgruppen unterteilt.

Die Anlagencodierung zeigt den Nachteil der nicht vereinheitlichten Datenquellen. Abhängig von der Datentabelle werden Attribute unterschiedliche codiert. Tabelle 25 zeigt den Umstand am Beispiel der Anlagencodierung. Ein ähnliches Bild zeigt sich bei anderen Attributen.

Tabelle 25: Codierungsbreite der Anlagen[534]

Short Name	ID	Engine	Long Name	Störcodetabelle
Compact 1	5	5	Compact 7326	7326
Compact 2	6	6	Compact 7327	6425
Compact 3	4	4	Compact 6427	6427
Dryer 1	7	2	Shrink 7325	7325
Dryer 2	2	1	Shrink 7324	7324
Stenter 1	3	3	Frame 6428	6428

Zur Zusammenführung musste über hinterlegte Übersetzungstabellen gearbeitet werden. Diese waren jedoch nicht so vollständig ausgeführt wie Tabelle 25, sondern jeweils mit der Gegenüberstellung von zwei Möglichkeiten. Der manuelle Aufwand war groß wie auch der Rechenaufwand im Rahmen einer möglichen Echtzeitausführung.

Innerhalb einer Tabelle konnte die einheitliche Darstellung und die eindeutige Auslegbarkeit nach den Formeln 6-2 und 6-4 jeweils mit „1“ bewertet werden. Nachdem zu den Tabellen Metadaten vorhanden waren die den Datentyp bzw. die Einheit beschrieben wäre Reifegrad „4“ möglich. Für die Analyse mussten die Daten jedoch in eine Gesamtdatenbasis zusammengeführt werden. Daher wurde der Reifegrad auf „2“ reduziert, da die einheitliche Darstellung nicht mehr gegeben ist.

Auf den ersten Blick war der angemessene Umfang durch die Aufzeichnungsdauer gegeben. Die Bewertung der Vollständigkeit würde ein anderes Bild ergeben, wenn alle Attribute einbezogen würden. Viele der Attribute waren nicht befüllt, da sie als Platzhalter dienten. Als Stichprobe wurden die Daten eines Monats ausgewertet. Nach Formel 6-7 kann die Vollständigkeit mit „1“ bewertet werden. Die Einträge ‚Null‘ oder ‚-1‘ sind keine Fehleinträge, sondern sind im Sinne der „informative missingness“ der Hinweis, dass kein Werkzeug gerüstet war, oder die Anlage ohne Auftrag gelaufen ist.

Die restlichen Attribute waren für die Analyse relevant und verwendbar. Eine visuelle Analyse zeigte, dass die diversen Attribute einen breit gefächerten Wertebereich hatten. Abbildung 46 zeigt die Störcodeverteilung eines der sechs Störungstypen einer der sechs Anlagen für zwei Jahre.

Der angemessene Umfang war gegeben, da Daten von drei Jahren vorlagen und diese bis auf Komponentenebene klassifiziert waren. Die Anzahl der Attribute war größer vier, was die Anwendung von komplexen Algorithmen für diagnostischer Analysen

[534] Quelle: Eigene Darstellung

rechtfertigte. In Summe kann die Reifegradkategorie Datenumfang mit vier bewertet werden.

Die Aktualität der Daten war hoch. Die Übertragung erfolgte über das MDE System unverzüglich in das IT-System. Der eingetragene Zeitstempel entsprach dem Zeitpunkt der Messung. Des Weiteren wurden für eine Messung mehrere Zeitstempel ermittelt. Für die Störung gab es jenen des Beginns und des Endes der Störung. Gleiches galt für andere Vorgänge und Bearbeitungsschritte, die erfasst wurden. Zusätzlich wurde die Konsistenz mit einer durchgehenden Auftragsnummer oder ID erhöht, mit welcher die Abarbeitung eines Auftrages verfolgt werden konnte. Dies ermöglicht Analysen zur Erstellung von Sequenzen. Die Reifegradkategorie Datenkonsistenz kann mit Reifegrad „4“ bewertet werden.

Es zeigte sich, dass die automatische Datenerfassung in der Operationalisierungsphase einen positiven Einfluss auf die restlichen Kategorien hat. Dieser positive Einfluss wurde bereits im House of Data Quality vermutet.

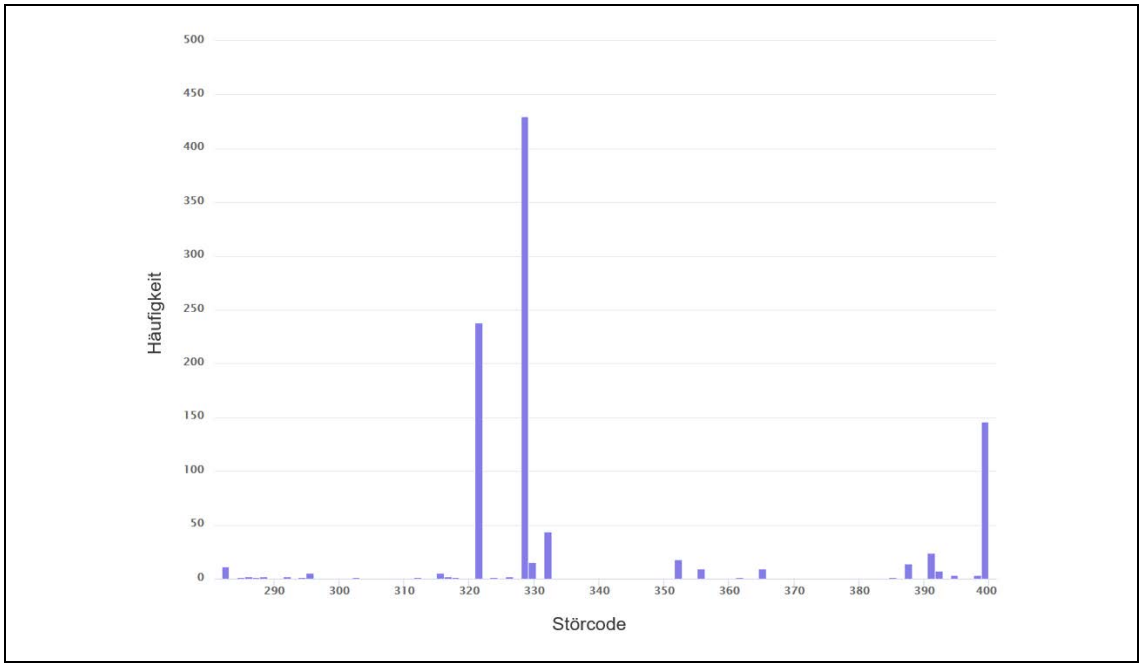

Abbildung 46: Störcodeverteilung[535]

Gesamtreifegrad:

Die Reifegradbewertung wird in Tabelle 26 zusammengefasst. Der Gesamtreifegrad der Organisation für datenanalytische Projekte ist mit „3“ zu bewerten. Für die Aufgabenstellung der diagnostischen Big Data gestützten Schwachstellenanalyse ist dieser Reifegrad ausreichend. Die Datenaufbereitung nimmt jedoch mehr Zeit als nötig wäre in Anspruch, da die Datenaufbereitung durch die Darstellung und Auslegung der Daten erschwert wird.

[535] Quelle: Eigene Darstellung

Tabelle 26: Reifegradbewertung Textilverarbeiter[536]

	Datenanalyse										
	Operationalisierung					Durchführung					
Reifegrad	Daten-erfassung	Datenbereit-stellung	Daten-formate			Daten-codierung	Daten-umfang	Daten-konsistenz			III
1											
2		X		X		X					
3	X		X						X		X
4							X	X			

Im konkreten Fall nahm das Datenverständnis und die Datenaufbereitung 25 Stunden in Anspruch, während die Modellierung fünf Stunden in Anspruch nahm. Die 25 Stunden hätten durch eine einheitlichere Darstellung und bessere Auslegbarkeit auf ca. zehn Stunden reduziert werden können.

Durchführungsphase

Um herauszufinden ob es eine Logik in der Eskalation von Fehlern gibt, wurde die Auswertung für die Störcodes durchgeführt. Das Ergebnis ist in Abbildung 47 zu sehen. Fehler „Err10“ gehört zu Störungstyp 1, dem unverzüglichen Stillstand der Anlage. Es ist daher nachvollziehbar, dass es sonst keine Störungen des Typs 1 gibt der diesen Störungstyp auslösen kann. Interessant sind Störungen des Störungstyps 6, die erst als Warnung eintreten. Diese können als Frühwarnindikatoren der Störung verwendet werden.

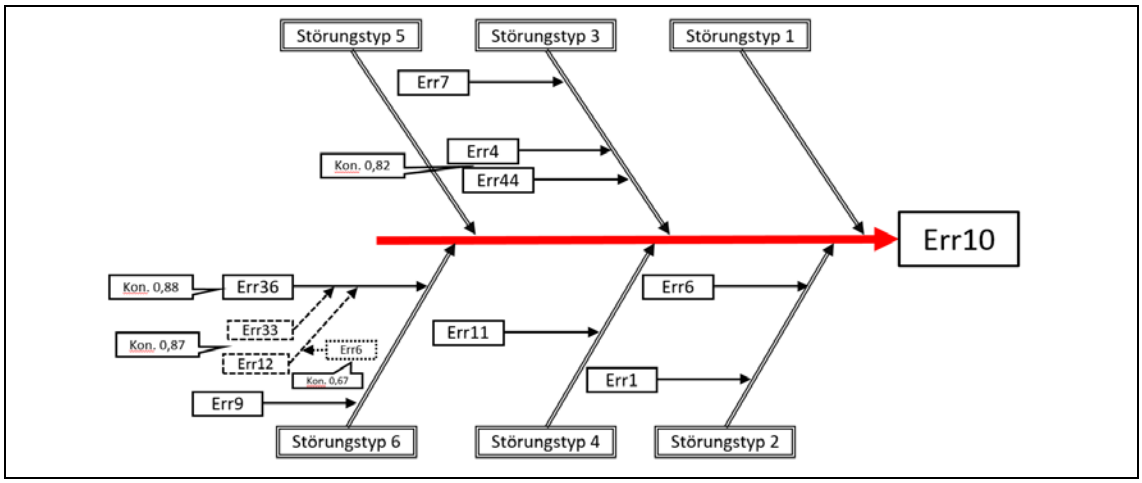

Abbildung 47: Big Data gestütztes Ishikawa Diagramm – Fehlercodes Textilverarbeiter[537]

[536] Quelle: Eigene Darstellung
[537] Quelle: in Anlehnung an Bernerstätter, R.; Kühnast, R. (2017), S. 180.

Die zweite Analyse wurde verwendet, um Zusammenhänge zu den Prozessparametern zu finden. Dazu mussten unterschiedliche Tabellen der Datenbanken verbunden werden. Durch die unterschiedlichen Darstellungen war dies mit erhöhtem Aufbereitungsaufwand verbunden. Für die Datenaufbereitung waren 25 Stunden nötig. Die interessanten Erkenntnisse waren der Zusammenhang zwischen einer Vorschubgeschwindigkeit und einer erhöhten Hitze von Trockenkammern. Abbildung 48 zeigt weitere Zusammenhänge.

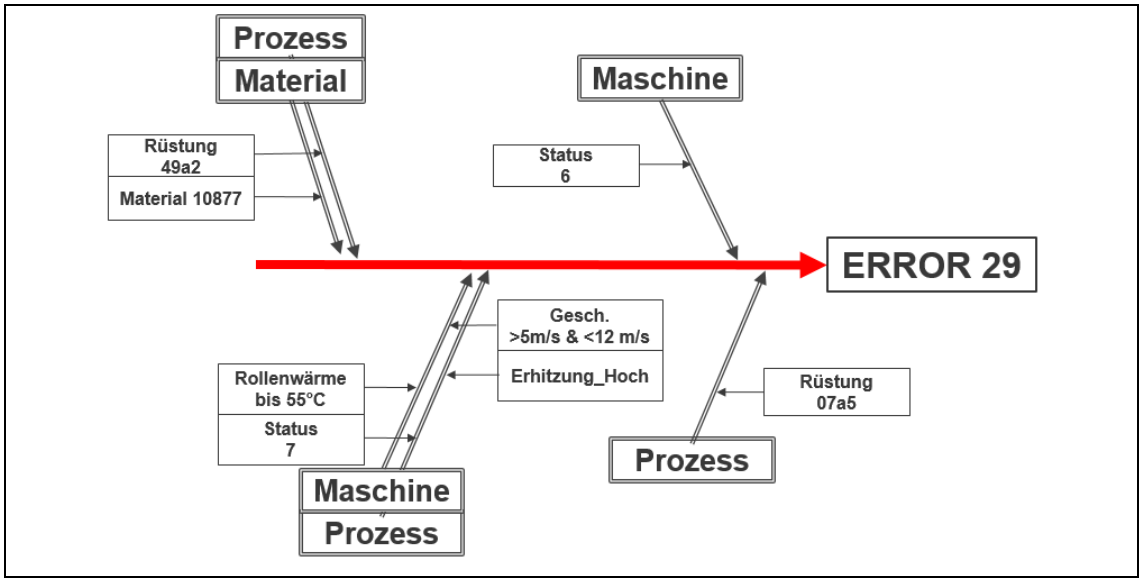

Abbildung 48: Big Data gestütztes Ishikawa Diagramm – Prozessparameter Textilverarbeiter[538]

Beitrag zum Reifegradmodell

Der Beitrag war eine weitere Abgrenzung der Reifegradkategorien, speziell in der Kategoire Datenkonsistenz. Erste Versuche zeigten vielversprechende Versuche der Big Data gestützten Schwachstellenanalyse mit Sequenzanalysen und Process-Mining.[539] Es war möglich den Fragenkatalog zu spezifizieren. Abbildung 49 zeigt das HoDQ wie es sich nach dem Fallbeispiel darstellt. Es ist vollständig und weitere Beispiele haben die Zusammenhänge weiter bestätigt. Es bestätigte sich die Vermutung, dass eine automatisierte Datenaufzeichnung einen sehr positiven Beitrag zum Reifegrad in den Kategorien Datenumfang und Datenkonsistenz bringt. Darüber hinaus veranschaulichte das Beispiel die Beeinflussung der Datendarstellung und der Datenkonsistenz durch die Datenbereitstellung. Die unterschiedlichen nicht integrierten Tabellen führten zu einer nicht einheitlichen Darstellung der Daten.

Die neue Art ein Ishikawa-Diagramm mithilfe von unterschiedlichen Prozessdaten qualitativ zu zeichnen ist der wesentliche Beitrag. Dieses Vorgehen kann als Metaartefakt des DSR in dieser Arbeit gesehen werden. Weitere Forschung dazu ist nötig.

[538] Quelle: Eigene Darstellung
[539] Siehe hierzu in Kapitel 8

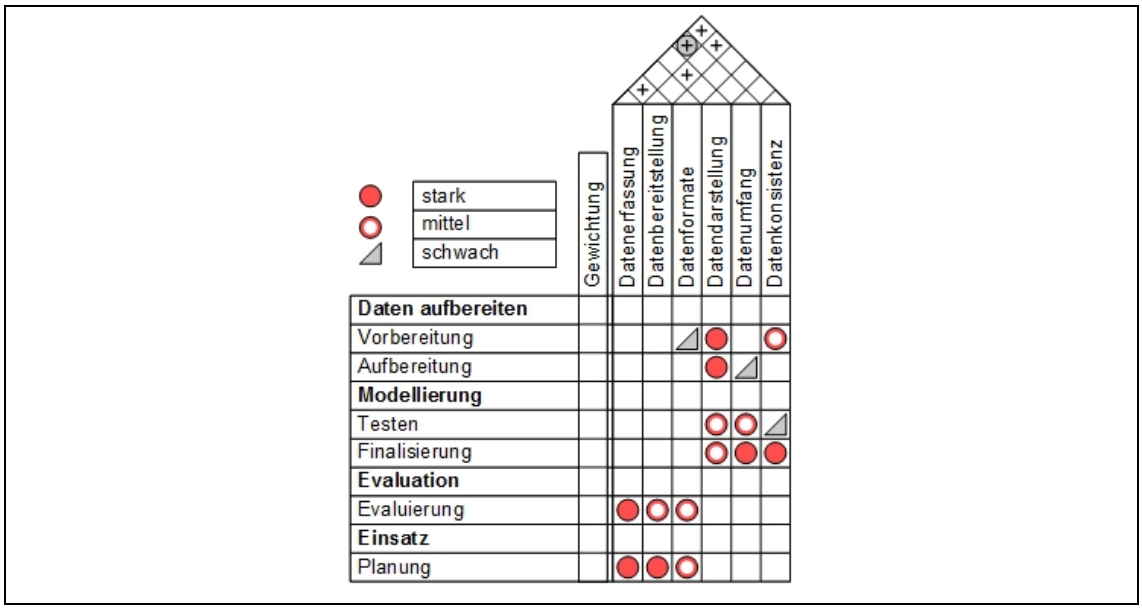

Abbildung 49: HoDQ durch Fallbeispiel Textilhersteller[540]

7.2.3 Fallbeispiel Holzverarbeiter

Das Unternehmen ist ein international tätiges österreichisches Familienunternehmen in der holzverarbeitenden Industrie. Im Rahmen einer Instandhaltungsreorganisation wurde als Teilprojekt eine Vorstudie zu den Möglichkeiten des Einsatzes von Data-Mining durchgeführt. Dieses Assessment wurde mit dem Reifegradmodell durchgeführt. Aufbauend auf den Ergebnissen wurde eine Big Data gestützte Schwachstellenanalyse durchgeführt. Im Rahmen des DSR wurde mit diesem Fallbeispiel das Reifegradmodell umfassend angewandt. Die Reifegradkategorien waren zu dieser Zeit fixiert, wie auch die Reifegradanzahl.

Vorbereitungsphase

Das Geschäftsziel, das durch den datenanalytischen Prozess erreicht werden soll war nicht klar definiert. Die Reifegraderhebung sollte ohne diese Inputs erfolgen und der Use Case auf dem Ergebnis basierend, definiert werden.

Um ein umfassendes Bild zu erhalten, wurden die Aufzeichnungen der Instandhaltung, der Prozesssteuerung und der Qualitätssteuerung ausgewertet. Die Analyse wird unter der Annahme einer später durchgeführten Schwachstellenanalyse beschrieben. Generell ist die Analyse ohne Fallbeispiel möglich.

Operationalisierung:

Die Datenaufzeichnung erfolgte unterschiedlich. Die Daten der Instandhaltungsdurchführung wurden manuell, nicht digital und unregelmäßig erfasst. Die Übertragung erfolgte am Schichtende in das IT-System. Die Glaubwürdigkeit und die Fehlerfreiheit sind mit „1“ zu bewerten. Der Reifegrad ist „1“.

[540] Quelle: Eigene Darstellung

Die Prozessdaten wurden vollautomatisch und regelmäßig alle zehn Sekunden aufgezeichnet. Ausnahmen bilden die Daten der Feuchtigkeitsmessung und Rezepteinstellungen die manuell eingegeben wurden. Das passierte direkt über den BDE Terminal. Die Glaubwürdigkeit der Erfassung ist sehr hoch, die Fehlerfreiheit ist hoch. Da händisch eingetragene Daten nachweislich falsch waren, ist der Gesamtreifegrad mit „3“ bewertet.

Die Qualitätsdaten wurden manuell, mit digitaler Unterstützung regelmäßig erfasst und direkt ins System übertragen. Glaubwürdigkeit und Fehlerfreiheit sind hoch (3). Der Reifegrad ist mit „3“ zu bewerten.

Insgesamt sind die Daten des Produktionsprozesses und der Qualitätssicherung aufgrund der Erfassung geeignet, um für komplexe Analyse verwendet zu werden, während die Instandhaltungsdaten das nicht sind. Bei einer Kombination aller Daten muss in Kauf genommen werden, dass die Daten der Instandhaltung die Ergebnisse verfälschen könnten, da die Aufzeichnungsqualität nicht sichergestellt ist. Dieses Faktum hat speziell in der Evaluierungsphase des CRIP-DM Bedeutung.

Die Daten der Instandhaltungsdurchführung wurden in SAP-PM gespeichert. Darauf kann eine Oracle-Datenbank zugreifen, um die Daten abzufragen. Die meisten Programme für Datenanalysen haben Schnittstellen zu Oracle. Der Reifegrad wird daher mit „3“ bewertet.

Die Daten des Prozesses werden in einem proprietären System gespeichert, zu dem es für externe Analyseprogramme keine Schnittstellen gibt. Eine Vereinheitlichung mit einem anderen System ist nur mit Exporten möglich und daher mit erheblichem Zeitaufwand verbunden. Die Zugänglichkeit wird mit „1“ bewertet. Es gibt keine weiteren positiven Einflüsse, weshalb der Reifegrad für die Prozessdaten in der Kategorie Datenhaltung mit „1“ bewertet wird.

Die Daten der Qualitätsteuerung wurden ebenfalls in einem proprietären System gehalten, zu dem es keine Schnittstellen nach außen gab. Der Datentransfer und die Erstellung von einem übergreifenden – horizontal wie vertikal integrierten – Datenbestand wäre nur mit Datenexporten und folglich viel Arbeitsaufwand möglich. Die Reife wird mit „1“ eingestuft.

Alle drei Datenquellen konnten Daten im XML Format exportieren. Der Reifegrad kann daher mit „2“ eingestuft werden.

Der Reifegrad der Einzelsysteme ist hetrogen. Die Datenaufzeichnung und die Datenformate sind von hoher Reife, mit Ausnahme der Aufzeichnung der Instandhaltungsdaten. Die Datenspeicherung ist jedoch nicht geeignet für systemübergreifendes Arbeiten. Die Empfehlung wäre eine Schnittstelle aller Systeme in ein Data-Warehouse. Die proprietären Systeme beliefern das Data-Warehouse, welches den Singel Point of Truth bildet. Data-Marts würden in weiterer Folge die Schnittstellen zu den Bereichen bilden und die relevanten Daten liefern. Mit OLAP werden erste Aufbereitungsschritte für komplexere Analysen durchgeführt.

Durchführung:

Die Instandhaltungsdaten waren nominal mit Standardcodes eingetragen und besitzen wo nötig für weitere Auskünfte Fließtext. Einige der Codes wurden in nicht atomarer

Form zusammengefasst. Die Darstellung wie die Auslegbarkeit sind gut, der Gesamtreifegrad wird mit „2“ festgelegt.

Das Skalenniveau der Prozessdaten ist abhängig vom Attribut unterschiedlich. Die Werte sind jedoch standardisiert und für die Daten liegen Metadaten vor, die die Einheit und die Skalierung definieren, sowie Auskunft über die Bedeutung geben. Die Darstellung ist einheitlich und die Auslegbarkeit eindeutig. Der Reifegrad kann für diese Daten in der Darstellungskategorie mit „4“ bewertet werden.

Die Qualitätsdaten sind metrisch bzw. nominal skaliert. Die Daten sind atomar abgespeichert. Dies gilt auch für den Zeitstempel bei dem Datum und Uhrzeit getrennt sind. Es gibt keine Metadaten zur genaueren Beschreibung, es sind jedoch die Attribute mit den passenden Einheiten aussagekräftig beschrieben. Der Reifegrad wird daher mit „4“ bewertet.

Die Instandhaltungsdaten umfassten zwei Aufzeichnungsjahre mit 7393 Einträgen über Stillstände mit sieben relevanten Attributen. Ein achtes Attribut welches zusätzlich den Störgrund genauer ausformuliert und in Fließtext codiert war, wird nicht betrachtet. Tabelle 27 zeigt die Bewertung der Einzelvollständigkeiten. In Summe ergibt sich eine Vollständigkeit von „4“ laut Formel 6-7.

Die Wertebereiche waren breit gestreut und können Relevanz haben. Ohne einen definierte Use-Case war die Feststellung der Relevanz nicht eindeutig durchführbar. Der angemessene Umfang war durch die Aufzeichnungsdauer und die Anzahl der Attribute gegeben. Der Reifegrad des Datenumfangs der Instandhaltungsdaten ist mit „4“ zu bewerten.

Tabelle 27: Vollständigkeit IH-Daten Holzverarbeiter[541]

Attribut	Fehlende Einträge	Einzel V laut Formel 6-7
Zeit 1	0	4
Zeit 2	0	4
Anlagenteil	0	4
Grund	6	0,999 = 4
Grundgruppe	58	0,99 = 4
Verursacher	0	1 = 4

Die Prozessdaten umfassten mehrere tausend Attribute mit ca. 130.000 Einträgen für drei Monate. Alle Daten waren vollständig erfasst und relevant, da sie einen großen Wertebereich besitzen. Der angemessene Umfang war wegen der geringen Speicherdauer von drei Monaten gering. Der Gesamtreifegrad wird daher mit „2“ bewertet.

Die Qualitätsdaten umfassten 1187 Attribute über einen Zeitraum von drei Jahren mit 2654 Einträgen. Es waren davon nur 15 vollständig erfasst. Von den nicht vollständig erfassten Attributen ist *V* laut Formel 6-6 kleiner 0,95. Die Vollständigkeit ist mit „1“ zu bewerten. Die Relevanz der 15 vollständig erfassten Daten ist hoch (3). Bei den

[541] Quelle: Eigene Darstellung

unvollständigen Daten ist diese mit nur einem Eintrag gering. Der angemessene Umfang kann aufgrund der Aufzeichnungsdauer und der Anzahl der Attribute als hoch (3) eingestuft werden.

Die geringe Vollständigkeit in Kombination mit der tlw. fehlenden Relevanz der Attribute, führte dazu, dass trotz der positiven Bewertung bei der Dimension angemessener Umfang, die Gesamtreife auf „2“ herabgestuft wurde, da für die Prognose von Stufe „3“ die Daten vollständiger aufgezeichnet werden müssten.

Die Konsistenz bei den Instandhaltungsdaten war aufgrund der händischen Aufzeichnung mit „1“ zu bewerten. Bei den Prozess- und Qualitätsdaten wurde die Reife mit „4“ bewertet. Der Zeitstempel beschrieb die abgespeicherten Prozessgrößen direkt. Die Übertragung erfolgte sofort und es wurden mehrere Zeitstempel eines Vorgangs erfasst.

Gesamtreifegrad:

Die Reifegradbewertung wird in Tabelle 28 zusammengefasst. Der Gesamtreifegrad der betrachteten Anlage ist mit „2“ zu bewerten. Die Einzelquellen haben in den Kategorien einen höheren Reifegrad. Wird jedoch angenommen, dass eine komplexere Analyse das Ziel wäre, mit der Absicht das Modell in den operativen Ablauf zu implementieren, ist der Gesamtreifegrad der Daten und des Datenmanagements dazu nicht in der Lage. Aus der Sicht der Reifegradanalyse sind diagnostische Analysen möglich. Für prognostische Analysen ist die Glaubwürdigkeit und die Fehlerfreiheit speziell der Instandhaltungsdaten nicht hoch genug. Aufgrund der Schwächen in der Datenhaltung wird eine Vereinigung der Daten für diagnostische Zwecke mit viel Arbeitsaufwand verbunden sein. Die definitive Handlungsempfehlung ist die Implementierung eines Data-Warehouses mit Data-Marts oder zumindest eine durchgehende Datenarchitektur in horizontaler Ebene.

Tabelle 28: Reifegradbewertung Holzverarbeiter[542]

		Reifegrad			
Kategorie	**Datenquellen**	**1**	**2**	**3**	**4**
Datenerfassung	IH	X			
	Prozess			X	
	Qualität			X	
	Gesamt		**X**		
Datenbereitstellung	IH			X	
	Prozess	X			
	Qualität	X			
	Gesamt	**X**			

[542] Quelle: Eigene Darstellung

Fortsetzung Tabelle 28: Reifegradbewertung Holzverarbeiter[543]

Datenformate	IH			X		
	Prozess			X		
	Qualität			X		
	Gesamt			**X**		
Operationalisierung	IH			X		
	Prozess			X		
	Qualität			X		
	Gesamt			**X**		
Datencodierung	IH			X		
	Prozess					X
	Qualität					X
	Gesamt				**X**	
Datenumfang	IH					X
	Prozess			X		
	Qualität			X		
	Gesamt			**X**		
Datenkonsistenz	IH		X			
	Prozess					X
	Qualität					X
	Gesamt			**X**		
Durchführung	IH		X			
	Prozess			X		
	Qualität				X	
	Gesamt			**X**		
Gesamt	IH		X			
	Prozess			X		
	Qualität				X	
	Gesamt			**X**		

Durchführungsphase

Mit dem vorliegenden Reifegrad wurde eine diagnostische Analyse als möglich erachtet. Um die Möglichkeiten zu beweisen, wurde vom Unternehmen eine Schwachstellenanalyse in Auftrag gegeben. Dem Unternehmen war die Schwachstelle und deren Ursache bereits bekannt. Ziel war es die Ursache mittel der Big Data gestützten Schwachstellenanalyse zu finden.

[543] Quelle: Eigene Darstellung

Die oben untersuchten Datenquellen wurden zur Bereitstellung eines gemeinsamen Datenbestandes herangezogen. Mittels einer Assoziationsanalyse wurden Regel abgeleitet und die relevanten in einem quantitativen Ishikawa-Diagramm dargestellt.

Erstes interessantes Ergebnis war, dass die gefundenen Regeln für die Ursache-Wirkungszusammenhänge produktrein waren. Auf der Anlage wurden zwei unterschiedliche Hauptprozessschritte ausgeführt. Die Daten beider waren in den drei betrachteten Quellen enthalten. Es kam nicht dazu, dass Variablen und Wertebereiche von Variablen eines Produktes in den Regeln des anderen Produktes vorkommen. Das spricht für die Analysemethode die speziell bei den Regeln, die durch die Interessantheitsmaße als besonders wichtig und aussagekräftig eingestuft wurden, keine Scheinzusammenhänge produziert.

Das zweite, für den Projektpartner interessante Ergebnis, war, dass die Analyse die Ursachen für den Ausfall gefunden hat. Häufige Ausfälle bei einem Austragsantrieb (IH-Daten) hatten die Ursache bei der Feuchte des Materials (Qualitätsdaten), der Prozessgeschwindigkeit (Prozessdaten), der Hallentemperatur (Prozessdaten) und einer Rezeptur (Produkt-/Prozessdaten). Die Regel

$$\{Vorschub[31{,}1-38{,}4]; Temp_{Halle}[31{,}5-40{,}6]; Feuchte_{Mat}[0-29]; Merkmal_{Produkt}[734]\} \rightarrow \{Ausfall\ Antrieb\}$$

hat eine Konfidenz von 100% und einen Lift von 15,23. Die Regel gilt somit immer, wenn die Kombination der Parameter im Regelrumpf eintritt. Der Lift besagt, dass der Ausfall um über 15-mal wahrscheinlicher bei der Parameterkombination eintritt, als bei anderen Parameterkombinationen in den Daten.

Abbildung 50 zeigt die Ursachen des Regelzusammenhangs für den Anlagenausfall als Ishikawa-Diagramm. Hier zeigt sich der Vorteil der Assoziationsanalyse. Anstatt mehrerer qualitativer Vermutungen deren Zusammenhang unbekannt sind, werden konkrete Ursachen aufgelistet, die über Regeln zusammenhängen.

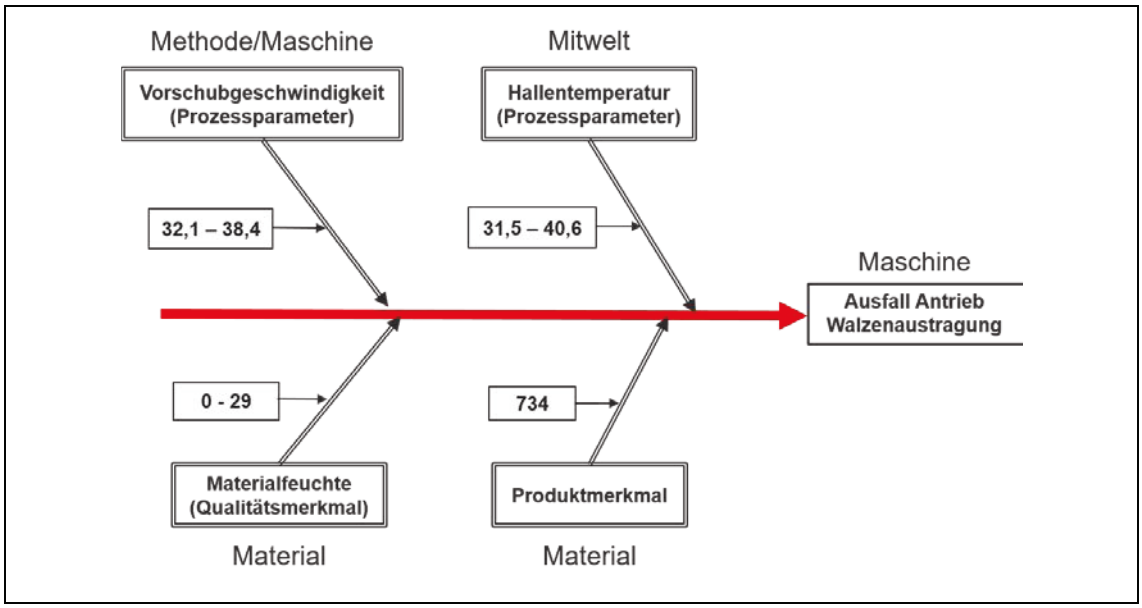

Abbildung 50: Big Data gestütztes Ishikawa Diagramm Holzverarbeiter

Laut Reifegradbewertung sollten sich die Daten gut für diagnostische Analysen eignen. Das Ergebnis bestätigt die Annahme dadurch, dass es sich bei dem Ausfall und den

Ursachen, um jene handelte, die der Projektpartner nach langer Suche selbst identifiziert hatte.

Operationalisierungsphase

Die Evaluierung der Regeln war durch die Darstellung im Ishikawa-Diagramm einfach, da das Format den Domänenexperten zugänglicher war, als die reine Niederschrift der Regel. Nachdem die Ursachen der Schwachstelle am Antrieb aus den Daten der Prozesssteuerung und der Qualitätssteuerung kamen, waren diese durch die Datenerfassung glaubwürdig und die Regeln galten als valide.

Der Einsatz des Vorgehens ist durch die Landschaft der Datenhaltungssysteme nicht automatisierbar und an Dritte nicht auslagerbar. Der geringe Reifegrad ist für den hohen Arbeitsaufwand verantwortlich, der in die Datenaufbereitung, speziell in die Vereinheitlichung der Daten, geflossen ist. Der Export der Daten aus den einzelnen Systemen als XML, erforderte es eine eigene Datenbank aufzusetzen, um die Dateien zu vereinheitlichen, und mit den IH-Daten als XLSX-Datei zu verknüpfen. In Summe flossen in diese Arbeiten 11,5 Personentage. Die Modellierung selbst benötigte nur 3,5 Personentage. Die zentrale Datenhaltung, hätte die Aufbereitung auf ca. 2-3 Personentage reduzieren können. Das entspricht der Arbeitszeit, die in diesem Projekt für die Aufbereitung nach der Vereinigung der Daten geflossen ist. Es zeigt sich, dass der geringe Reifegrad in der Operationalisierung des CRISP-DM den Arbeitsaufwand in der Durchführungsphase stark erhöht. Da durch die vielen Schnittstellen und Zwischenschritte die Rechenzeit zu groß wird, eignen sich Reifegrade unter „3“ für Echtzeitanalysen, wie Prognosen, nicht.

7.3 Zusammenfassung und Reflexion der Fallbeispiele

Die sechs Fallbeispiele gaben den Input für die kontinuierliche Entwicklung des Reifegradmodells durch den DSR Zyklus. Die drei ersten Fallbeispiele definierten den Bedarf, die grobe Struktur des Modells und den Inhalt der Reifegradkategorien und Reifegrade. Im Sinne des DSR wurden nach jedem Projekt die Erkenntnisse verwendet, um das aktuelle Modell neu zu evaluieren und weiter zu verfeinern. Die ersten beiden übernahmen darin eine besondere Rolle, da sie durch die Projekterfahrung zeigten, wo die wesentlichen Probleme bei der Umsetzung waren. Die Erfassungsmethoden, Speichermöglichkeiten, Datenformate usw. lieferten wichtige Referenzen zur weiteren Recherche und Abgrenzung der Reifegrade. Der Stahlverarbeiter konnte bereits bewertet werden, da die Extrema der Kategorien bekannt waren[544]. Hier zeigte sich die Problematik der unterschiedlichen Datenquellen deutlich. Während die Prozessdaten sehr gut in der Eignung für Prognosen waren, war deren Zugang problematisch. Die Instandhaltungsdaten verschärften das Problem durch die schlechte Erfassung. Das Gesamtsystem war in Folge für Prognosen ungeeignet. Die unabhängigen Variablen, jene der Sensormessungen, wurden gut erfasst, die abhängigen (die zu prognostizierenden) waren von schlechter Qualität. Daher nur Reifegrad „2“ in Summe.

[544] Vgl. Bernerstätter, R. (2018b), S. 33.

Das Fallbeispiel des Automobilherstellers bildet eine interessante Zwischenstufe. Das Reifegradmodell war bereits ausgereift genug, um zielgerichtete Aussagen zu treffen. Es wurde erkannt, dass eine Prognose mit den vorhandenen Daten nicht möglich war. Das Problem war ähnlich jenem des Stahlverarbeiters. Die Qualität der Daten der abhängigen Variablen war nicht ausreichend. Es wurde eine Anomalieerkennung implementiert, die im weiteren Sinn als diagnostische Analyse betrachtet werden konnte. „Warum ist etwas passiert?“ ist die zentrale Frage. In der Definition der Schwachstelle aus Abschnitt 5.3 ist die Anomalie eine Schwachstelle, da sie eine negative Wirkung ist, die an der Messstelle auftritt. Die Ursache ist zeitlich betrachtet vorgelagert und nicht direkt an die Messstelle gebunden. Das Ereignis ist die Anomalie, die Ursache ist in den Daten zu finden die zur Zeit der Anomalie anliegen. Die Anomalieerkennung konnte erfolgreich implementiert werden. Die Einordnung von Gesamtreifegrad „2“ war folgerichtig korrekt.

Die letzten drei Beispiele dienten zur konkreteren Anwendung anhand der Schwachstellenanalyse. Die Analysen, die Fragen und die Ausgestaltung des Reifegradmodells haben den Anspruch allgemeingültige Erkenntnisse zu liefern, die eine generelle Einordnung eines Beispiels ermöglichten. Die Schwachstellenanalyse war dafür ein wichtiger Ankerpunkt.

Diagnostische Erkenntnisse sind wichtige Zwischenschritte für die Weiterentwicklung zur prognostischen und präskriptiven Analyse.[545] Die Abgrenzung des Reifegrades „2“ legt das Fundament für die Folgenden und definiert den schlechteren.

Der Hausgerätehersteller zeigte die Möglichkeiten der Bewertung des Reifegradmodells von Anwendungsfälle, die nicht der Komplexität von Industrial Analytics entsprechen. Mit diesem Fallbeispiel erfolgte die Abgrenzung des diagnostischen zum deskriptiven Reifegrad.

Der Anwendungsfall des Textilherstellers verdeutlichte den Zusammenhang zwischen Datenerfassung, dem Umfang und der Konsistenz der Daten zum einen. Zum anderen zeigte sich der Zusammenhang zwischen der Datenbereitstellung mit der Schnittstellenproblematik und des damit verbundenen steigenden Arbeitsaufwandes in der Datenaufbereitung. Des Weiteren kamen die Assoziationsanalyse und das Ishikawa-Diagramm in der neuen Form zum Einsatz.

Die Analyse des Reifegrades beim Holzverarbeiter zeigte die gute Anwendbarkeit des Reifegradmodells. Einerseits war im ersten Schritt kein Use-Case bekannt. Trotzdem konnte der Reifegrad und die Möglichkeiten ermittelt werden. Andererseits zeigte sich, dass das Modell von anderen Personen angewandt werden konnte. Das Vorgehensmodell ist universell und strukturiert, sodass es von Data-Mining Experten bzw. in der Datenanalytik erfahrenen Personen angewandt werden kann. Das erfüllt das pragmatische Merkmal eines Modells, das besagt, dass das Modell von Personen eingesetzt werden können soll, die die Kompetenz dafür haben.

Mit Ausnahme des Fallbeispiels des Hausgeräteherstellers ermöglichten die restlichen Fünf eine grobe Abschätzung des zusätzlichen Arbeitsaufwandes durch die fehlende Reife in den verschiedenen Kategorien. Tabelle 29 gibt darüber eine Übersicht. Die

[545] Vgl. Karim, R. et al. (2016), S. 218.

Kategorien der Operationalisierungsphase haben den größten Einfluss auf den vermeidbaren Arbeitsaufwand. Dieser entsteht hauptsächlich durch fehlende Schnittstellen und Standards in den Formaten und der Darstellung der Daten.

Tabelle 29: Geschätzter vermeidbarer Arbeitsaufwand durch höhere Reife[546]

Fallbeispiel	Geschätzter vermeidbarer Aufwand	Grund

Automobilproduzent	ca. 1 Personenmonat	• Datenbereitstellung
		  ○ Proprietäres System – fehlende Schnittstellen
		• Datendarstellung
		  ○ Fehlendes Codierungsschema

Stahlverarbeiter	ca. 3 Personentage	• Datenbereitstellung ○ Proprietäres System – fehlende Schnittstellen

Automobilzulieferer	ca. 1,5 Personen-wochen	• Datenerfassung
		  ○ Unregelmäßige Erfassung
		• Datenkonsistenz
		  ○ Offset

Textilhersteller	ca. 25 h	• Datendarstellung ○ Nicht einheitliche Darstellung

Holzverarbeiter	ca. 7 Personentage	• Datenbereitstellung und Datenformate
		  ○ Nicht integrierte Systeme
		  ○ XML-Daten in DB-System übertragen

[546] Quelle: Eigene Darstellung

8 Zusammenfassung und Ausblick

In den folgenden Abschnitten werden die wesentlichen Punkte der Arbeit zusammengefasst. Die Ergebnisse werden kritisch reflektiert und auf die Beantwortung der Forschungsfragen explizit hingewiesen. Zum Abschluss wird der weitere Forschungsbedarf dargelegt.

8.1 Zusammenfassung

Die zunehmende Digitalisierung und die damit entstehenden Datenmengen bringen Unternehmen vermehrt dazu Erfahrungen mit der Analyse der Daten zu sammeln. Die gewonnenen Erkenntnisse dienen nicht nur der Entwicklung neuer Geschäftsmodelle und Dienstleistungen, sondern auch der Identifizierung von Verbesserungspotenzialen in den unternehmenseigenen Prozessen.

Die Unternehmen stehen vor der Herausforderung den Anwendungsfall für die Datenanalysen genau zu definieren. Sie können diesen an ihre eigenen Fähigkeiten anpassen; idealerweise entwickeln sie sich jedoch selbst weiter, um die gewünschte Aufgabenstellung zu erfüllen. Die Einschätzung der eigenen Fähigkeiten ist ein zentrales Problem, da die Analyseprozesse von Data-Mining Projekten abstrakte, bisher unbekannte Anforderungen an das Unternehmen stellen. Daher ist es nötig vor dem Start eines solchen Projektes jenen IST-Zustand des Unternehmens abzubilden, der sich auf die Inputfaktoren solcher analytischen Anwendungen konzentriert. Dafür wurde in dieser Arbeit ein Reifegradmodell entwickelt, das diesen IST-Zustand erhebt. Das Modell sollte generisch auf alle datenanalytischen Anwendungen anwendbar sein und sich speziell auf die Daten als zentraler Inputfaktor in der Datenanalytik beziehen. Der Grund liegt in der Datenqualität als immer wieder kehrender Faktor bei gescheiterten Analyseprojekten. Da sich dieses Modell auf die Bewertung von unternehmensinternen Projekten konzentriert und nicht auf die Möglichkeit zur Ableitung neuer digitaler Geschäftsmodelle, wurde die Schwachstellenanalyse als exemplarischer Prozess gewählt, um den effizienzsteigernden Verbesserungscharakter abzubilden. Das Modell soll die Unternehmen dabei unterstützen sich weiterzuentwickeln und Handlungspotenziale aufzeigen. Dadurch ist es ihnen möglich - wie oben beschrieben - die eigenen Fähigkeiten den Anforderungen des Analyseprozesses anzupassen.

Es besteht der Anspruch, dass das Modell für unterschiedliche Analysearten einsetzbar ist. Dazu muss es in der Bewertungshierarchie entsprechend der Komplexitäten der Analyseprozesse aufgebaut werden.

Für die Erarbeitung der Anforderungen wurde die Literatur der Reifegradmodelle, der Daten- und Informationsqualität und der datenanalytischen Grundlagen aufgearbeitet und kritisch gewürdigt. Die Schwachstellenanalyse gab den Rahmen für die Entwicklung vor. Um die generelle Einsetzbarkeit über verschiedene datenanalytische Fallbeispiele hinaus zu gewährleisten, wurde der CRISP-DM Prozess für Data-Mining als generischer Prozess identifiziert. Aufbauend auf diesen wurde das Reifegradmodell entwickelt.

Die Datenqualität wurde als Ebene in das Reifegradmodell aufgenommen. Die bereits existierenden Betrachtungen geben einen wichtigen Input in die Ausgestaltung der Reifegradkategorien und setzen ein theoretisch argumentiertes Mindestmaß an Punkten, die in der Ausformulierung der Reifegradkategorien betrachtet werden muss. Die Betrachtung des Daten- und Informationsmanagements und der datenanalytischen Grundlagen schließt die Lücke, die durch die klassischen Datenqualitätsdimensionen nicht abgedeckt wird. Darauf aufbauend wurden sechs Reifegradstufen abgeleitet, die den Datenqualitätsdimensionen und den Prozessschritten des CRISP-DM zugeordnet werden konnten.

Die Hierarchie der Reifegradstufen wurde ebenfalls aus den datenanalytischen Grundlagen abgeleitet. Die Analysekonzepte werden anhand ihrer Komplexität gereiht. Je komplexer die Analyse, desto höher der Unternehmensnutzen und desto höher sind die Reifegradansprüche an die einzelnen Kategorien. Abgeleitet aus der Aufgabenstellung des Unternehmens, werden Analysemethoden definiert, die wiederum den Komplexitätsstufen der Analysekonzepte zugeordnet sind. Damit ist es möglich den SOLL-Zustand auf der Hierarchiestufe der Reifegrade zu definieren. Handlungsempfehlungen zeigen den Unternehmen Potenziale zur Entwicklung auf einen höheren Reifegrad. Die Empfehlungen können in einem Projektablauf berücksichtigt werden, bevor mit der Datenanalyse begonnen wird. Das soll zu realistischen Zeit- und Kostenplänen führen und bei Umsetzung der Maßnahmen den Erfolg des datenanalytischen Projektes gewährleisten. Unterstützt wird die Maßnahmenableitung durch einen Fragebogen, der auf die Reifegradkategorien und die Reifegrade abgestimmt ist. Durch die Antworten, bzw. die fehlenden Antworten, werden die Lücken zu den nächsten Reifegraden bestimmt und die Handlungsempfehlungen abgeleitet.

Um den Prozess der Schwachstellenanalyse bei großen Datenmengen zu unterstützen wurde mit der Assoziationsanalyse eine geeignete datenanalytische Methode identifiziert. Sie bildet die methodische Referenz bei der Ausgestaltung der Kategorien. Darüber hinaus fand sie exemplarische Anwendung in Fallbeispielen, um das Reifegradmodell laufend zu verbessern. Die Ergebnisse der Analysen wurden in einem abgewandelten Ishikawa-Diagramm visualisiert.

Die Entwicklung des Reifegradmodells wurde im Rahmen von Design Science Research durchgeführt. Dieser sieht die Erstellung des Hauptartefakts, hier des Reifegradmodells, anhand von praktischen Fallbeispielen vor. Für diese Arbeit wurden sechs Fallbeispiele erörtert. Drei davon gingen über die Schwachstellenanalyse hinaus, um die generelle Einsetzbarkeit zu gewährleisten und drei beschäftigten sich mit der Schwachstellenanalyse. Die Anwendbarkeit des Reifegradmodells und der Assoziationsanalyse mit dem datengestützten Ishikawa-Diagramm wurde in den Fallbeispielen bestätigt.

8.2 Kritische Würdigung

Das Reifegradmodell hat den CRISP-DM Zyklus und die Schwachstellenanalyse als starkes Fundament für die Entwicklung eingesetzt. Speziell der CRISP-DM ist ein generisch einsetzbarer Analyseprozess, der in einer breiten Palette von datenanalytischen Anwendungen genutzt wird. Es kann keineswegs ausgeschlossen werden, dass die Anwendung des Reifegradmodells in Projekten, die nicht auf den CRISP-DM basieren problemlos möglich ist.

Außerdem soll nicht unerwähnt bleiben, dass die vier Reifegrade aus der Literatur abgeleitet und nicht für jeden Beispiele umgesetzt wurden. Reifegradstufe eins und zwei wurden in Fallbeispielen getestet. Reifegradstufe drei wurde zum Teil in einem Fallbeispiel validiert, wobei die zentrale Aussage war, dass die Umsetzung des Anwendungsfalles auf dieser Ebene nicht problemlos ist. Reifegradstufe vier wurde induktiv abgeleitet und basiert auf der logischen Entwicklung über Reifegradstufe drei hinaus. Eine Validierung konnte aufgrund fehlender Fallbeispiele nicht durchgeführt werden.

Auf die Beantwortung der Hauptforschungsfrage wurde in der Arbeit durchgehend hingearbeitet. Sie endet in Kapitel 6 mit der Entwicklung des Reifegradmodells und in Kapitel 7 mit dessen Validierung.

Die Subforschungsfragen wurden in folgenden Kapiteln und Abschnitten beantwortet:

1. „Mit welchen Reifegradkategorien und Reifegradstufen lassen sich diese Faktoren und Komplexitätsgrade beschreiben?“ *in Abschnitt 6.3.1 und 6.4.1*
2. „Lassen sich bestehende Datenqualitätsbetrachtungen in die Bewertungslogik integrieren bzw. was muss dafür verändert werden?“ *in Abschnitt 3.4 und 6.2.3*
3. „Sind Zusammenhänge zwischen der Reife des Datenmanagements und der Datenqualität feststellbar?“ *in Abschnitt 6.4.2 und Kapitel 7*
4. „Wie kann die Schwachstellenanalyse in die Logik der Analysekomplexität des Data-Minings integriert werden?“ *in Abschnitt 4.2.1 und 4.3*
5. „Wie müssen die Reifegrade für die Schwachstellenanalyse beschrieben werden und wie kann der Reifegrad für diese erhoben werden?“ *in Abschnitt 6.4.1, 6.4.2 und 7.2. und Anhang 1A*
6. „Welche datenanalytischen Methoden können die Schwachstellenanalyse bei der steigenden Systemkomplexität unterstützen?“ *in Abschnitt 4.2.2, 7.2.2 und 7.2.3*

8.3 Ausblick und weiterer Forschungsbedarf

Das Reifegradmodell nimmt keine monetäre Bewertung der Reifegradverbesserung vor. Die Definition von quantifizierbaren Informationseinheiten zur Bestimmung eines Grenznutzens von Information ist ein Forschungsfeld, welches dabei unterstützen könnte, die Handlungsempfehlungen, die sich aus dem Reifegradmodell ergeben, wirtschaftlich zielgerichteter abzugeben. Ansätze zur Kostenbewertung finden sich in HEIL, der die Kostenwirkung von Störungen betrachtete.[547] Nachdem das

[547] Vgl. Heil, M. (1995), S. 100 ff.

Reifegradmodell mit dem Fokus der Schwachstellenanalyse entwickelt wurde kann an dieser Stelle die Forschung fortgesetzt werden.

Der Zusammenhang zwischen dem Datenmanagement, der Datenqualität und folglich den Einsatzmöglichkeiten für Industrie 4.0 Anwendungen wurde in Ansätzen festgestellt. Die vertiefende Betrachtung der durchgehenden und standardisierten Architektur von Datenmanagementsystemen und die Modellierung von Daten auf die Anwendungsmöglichkeiten sind weitere Forschungsgebiete. Es besteht die Möglichkeit, dass die Reifegradbetrachtung infolge dieser Forschungsergebnisse nur mehr auf das Datenmanagement fokussiert werden muss, da die Daten unweigerlich eine hohe Qualität aufweisen.

Literaturverzeichnis

Abbott, D. (2014): Applied predictive analytics: principles and techniques for the professional data analyst. Indianapolis: Wiley. ISBN 978-1-118-72796-6.

Adam, D. (1993): Planung und Entscheidung: Modelle - Ziele - Methoden ; mit Fallstudien und Lösungen. 3., vollst. überarb. und erw. Aufl, Wiesbaden: Gabler. ISBN 978-3-409-34613-9.

Ahlemann, F.; Schroeder, C.; Teuteberg, F. (2005): Kompetenz- und Reifegradmodelle für das Projektmanagement: Grundlagen, Vergleich und Einsatz. Osnabrück: Univ., FB Wirtschaftswiss., Organisation u. Wirtschaftsinformatik. ISBN 978-3-936475-24-1.

Apel, D.; Behme, W.; Eberlein, R.; Merighi, C. (2015): Datenqualität erfolgreich steuern - Praxislösungen für Business-Intelligence-Projekte. 3., überarb. und erw. Aufl, Heidelberg: dpunkt. ISBN 978-3-86490-042-6.

Augustin, S. (1990): Information als Wettbewerbsfaktor: Informationslogistik - Herausforderung an das Management. Zürich: Verl. Industrielle Organisation [u.a.]. ISBN 978-3-85743-949-0.

Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (2016): Multivariate Analysemethoden: eine anwendungsorientierte Einführung. 14., überarbeitete und aktualisierte Auflage, Berlin Heidelberg: Springer Gabler. ISBN 978-3-662-46075-7.

Bange, C.; Janoschek, N. (2014): Big Data Analytics - Auf dem Weg zur datengetriebenen Wirtschaft. Würzburg: BARC GmbH. 2014.

Baumöl, U.; Meschke, M. (2009): Das Management von Datenqualität: Transfer. In: Controlling & Management, Jg. 53, Nr. 1, S. 62–65.

Becker, J.; Knackstedt, R.; Pöppelbuß, J. (2009): Entwicklung von Reifegradmodellen für das IT-Management: Vorgehensmodell und praktische Anwendung. In: WIRTSCHAFTSINFORMATIK, Jg. 51, Nr. 3, S. 249–260.

Beekmann, F. (2003): Stichprobenbasierte Assoziationsanalyse im Rahmen des Knowledge Discovery in Databases. 1. Aufl, Wiesbaden: Dt. Univ.-Verl. ISBN 978-3-8244-2168-8.

Behrenbeck, K. R. (1994): DV-Einsatz in der Instandhaltung: Erfolgsfaktoren und betriebswirtschaftliche Gesamtkonzeption. Wiesbaden: Dt. Univ.-Verl. [u.a.]. ISBN 978-3-8244-6076-2.

Bennett, N.; Lemoine, G. J. (2014): What a difference a word makes: Understanding threats to performance in a VUCA world. In: Business Horizons, Jg. 57, Nr. 3, S. 311–317.

Bernerstätter, R. (2018a): Big Data Analytics, Predictive Maintenance und Schwachstellenanalyse. In: Biedermann, H. (Hrsg.): Benchmark -

Instandhaltung; Eine Studie zum Reifegrad von Instandhaltungsorganisationen der DACH-Region. Köln: TÜV Media. , S. 135–139.

Bernerstätter, R. (2018b): Data Maturity for Smart Factory Applications – An Assessment Model. In: ACTA TECHNICA CORVINIENSIS – Bulletin of Engineering, Jg. 11, Nr. 1, S. 31–35.

Bernerstätter, R. (2018c): Daten als Ressource in Industrie 4.0: Kosten und Nutzen von Datenqualität. In: WINGbusiness, Jg. 2018, Nr. 1, S. 6–39.

Bernerstätter, R.; Hirschmugl, R. (2018): Anomalieerkennung an Altanlagen durch minimale Hardwarenachrüstung und Data Analytics. In: Predictive Maintenance - Realität und Vision. Köln: TÜV Media. (Praxiswissen Instandhaltung)., S. 201–213.

Bernerstätter, R.; Kühnast, R. (2017): Schwachstellenanalyse zur Gewährleistung der Handlungsfähigkeit in komplexen Systemen. In: Biedermann, H. (Hrsg.): Erfolg durch Lean Smart Maintenance - Bausteine und Wege des Wandels. Köln: TÜV Media. (Praxiswissen Instandhaltung)., S. 163–183.

Bernerstätter, R.; Kühnast, R. (2018): Data Maturity Assessment – Bewertung der Reife des Datenmanagements für Smart Maintenance. In: BHM Berg- und Hüttenmännische Monatshefte.

Bernerstätter, R.; Nemeth, T.; Glawar, R.; Habersohn, C.; Biedermann, H. (2016): Instandhaltung 4.0 - Sicherung der Produktqualität und Anlagenverfügbarkeit durch einen echtzeitbasierten Instandhaltungsleitstand. In: WINGbusiness, Jg. 2016, Nr. 01, S. 25–28.

Biedermann, H. (1985): Optimale Instandhaltungspolitik durch Kenntnis der Ausfallkosten unter besonderer Berücksichtigung der Hüttenindustrie. In: BHM Berg- und Hüttenmännische Monatshefte, Jg. 130, Nr. 7, S. 223–229.

Biedermann, H. (1988): Instandhaltungs-Controlling mittels Kennzahlen-Kennzahlen als Führungsinstrument für Analyse, Planung, Steuerung und Kontrolle. In: Männel, W.; Aman, K. (Hrsg.): Integrierte Anlagenwirtschaft. Köln: Verl. TÜV Rheinland. ISBN 978-3-88585-467-8 (Schriftenreihe Anlagenwirtschaft)., S. 305–329.

Biedermann, H. (2008): Anlagenmanagement: Managementinstrumente zur Wertsteigerung. 2., vollst. überarb. und aktualisierte Aufl, Köln: TÜV Media. ISBN 978-3-8249-1080-9.

Biedermann, H. (2016): Lean Smart Maintenance - Wertschöpfende, lernorientierte und ressourceneffiziente Instandhaltung. In: Biedermann, H. (Hrsg.): Lean Smart Maintenance: Konzepte, Instrumente und Anwendungen für eine effiziente und intelligente Instandhaltung. Köln: TÜV Media. (Praxiswissen für Ingenieure - Instandhaltung)., S. 19–29.

Biedermann, H. (2017): Lean Smart Maintenance - Controlling. Die Schwachstellenanalyse als zentrales Element im Führungssystem der Instandhaltung. In: Biedermann, H. (Hrsg.): Erfolg durch Lean Smart Maintenance: Bausteine und Wege des Wandels. Köln: TÜV Media GmbH. ISBN 978-3-7406-0243-7 (Praxiswissen für Ingenieure - Instandhaltung)., S. 23–36.

Biedermann, H. (2018): Predictive Maintenance-Möglichkeiten und Grenzen. In: Biedermann, H. (Hrsg.): Predictive Maintenance; Realität und Vision. Köln: TÜV Media. ISBN 978-3-7406-0359-5 (Praxiswissen Instandhaltung)., S. 23–40.

Biedermann, H.; Bernerstätter, R.; Kinz, A. (2017): Lean Smart Maintenance - Ressourceneffiziente, wertschöpfungs- und lernorientierte Instandhaltung. In: Horn, G. (Hrsg.): Der Instandhaltungs-Berater. 73. Ergänzung, Köln: TÜV Media. , S. 56.

Bodendorf, F. (2006): Daten- und Wissensmanagement. 2., aktualisierte und erweiterte Auflage, Berlin Heidelberg New York: Springer. ISBN 978-3-540-28743-8.

Bogaschewsky, R.; Müller, H. (2016): Industrie 4.0: Wie verändern sich die IT-Systeme in Einkauf und SCM? Bundesverband Materialwirtschaft, Eunkauf und Logistik. 2016.

Bollinger, T. (1996): Assoziationsregeln - Analyse eines Data Mining Verfahrens. In: Informatik-Spektrum, Jg. 19, Nr. 5, S. 257–261.

Brin, S.; Motwani, R.; Ullman, J. D.; Tsur, S. (1997): Dynamic itemset counting and implication rules for market basket data. Proceedings of the 1997 ACM SIGMOD international conference, Tucson, Arizona, United States, 1997. Tucson, Arizona, United States: ACM Press. ISBN 978-0-89791-911-1.

Brunner, M.; Jodlbauer, H.; Schagerl, M. (2016): Reifegradmodell Industrie 4.0; Unternehmen durch Industrie 4.0 stärken. In: Industrie 4.0 Management, Jg. 32, Nr. 5, S. 49–52.

Brust, O.-E.; Möller, F.; Skrablies, W. (2015): Informationsqualität für das Management mit TOPAS®. In: Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. (Hrsg.): Daten- und Informationsqualität. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN 978-3-658-09213-9, S. 359–378.

Budde, L.; Friedli, T. (2017): Komplexitätsmanagement in Zeiten von Industrie 4.0 und wachsender Digitalisierung. In: Wirtschaftsinformatik & Management, Jg. 9, Nr. 2, S. 28–39.

Camm, J. D.; Cochrane, J. J.; Fry, M. J.; Ohlmann, J. W.; Anderson, D. R.; Sweeney, D. J.; Williams, T. A. (2018): Business analytics; Descriptive, Predicitve, Prescriptive. 3rd edition, Boston: Cengage Learning. ISBN 978-1-337-40642-0.

Chamoni, P.; Gluchowski, P. (2006): Analytische Informationssysteme - Einordnung und Überblick. In: Chamoni, P. (Hrsg.): Analytische Informationssysteme: Business-Intelligence-Technologien und -Anwendungen ; mit 13 Tabellen. 3., vollst. überarb. Aufl, Berlin: Springer. ISBN 978-3-540-29286-9, S. 3–22.

Chandola, V.; Banerjee, A.; Kumar, V. (2009): Anomaly detection: A survey. In: ACM Computing Surveys, Jg. 41, Nr. 3, S. 1–58.

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (2000): CRISP-DM 1.0; Step-by-step data mining guide, SPCC Inc. 2000.

Chrissis, M. B.; Konrad, M.; Shrum, S. (2003): CMMI: guidelines for process integration and product improvement. Boston: Addison-Wesley. ISBN 978-0-321-15496-5.

Christmann, A. (1988): Kommunales Informationsmanagement (Teil 1): Ein Weg aus dem Dilemma?, ÖVD/online. 1988.

Cios, K. J.; Kurgan, Lukasz A. (2004): Trends in Data Mining and Knowledge Discovery. In: Pal, N. R.; Jain, L. C. (Hrsg.): Advanced techniques in knowledge discovery and data mining. New York: Springer-Verlag. ISBN 978-1-85233-867-1 (Advanced information and knowledge processing)., S. 1–26.

Cios, K. J.; Pedrycz, W.; Swiniarski, R. W.; Kurgan, L. A. (2007): Data mining: a knowledge discovery approach. New York, NY: Springer. ISBN 978-0-387-33333-5.

Cleve, J.; Lämmel, U. (2014): Data mining. Müchen: De Gruyter Oldenbourg. ISBN 978-3-486-71391-6.

CMMI Product Team (2010): CMMI® for Acquisition, Version 1.3 - Improving processes for acquiring better products and services. Carnegie Mellon University: Software Engineering Institute. Software Engineering Process Management Program. 2010.

Cronholm, S.; Göbel, H. (2016): Evaluation of the Information Systems Research Framework; Empirical Evidence from a Design Science Research Project. In: The Electronic Journal Information Systems Evaluation, Jg. 19, Nr. 3, S. 158–168.

Crosby, P. B. (1980): Quality is free: the art of making quality certain. New York; Scarborough (Ontario: New American Library. ISBN 978-0-451-62129-0.

Crosby, P. B. (2000): Qualitätsmanagement. Sonderausg., aktualisierte Auflage des Weltbestsellers ”Quality is free”, Wien: Ueberreuter. ISBN 978-3-7064-0643-7.

Dean, J. (2014): Big data, data mining, and machine learning: value creation for business leaders and practitioners. Hoboken, NJ: Wiley. ISBN 978-1-118-92070-1.

Deeskow, P.; Steinmetz, U.; Hay, M. (2008): Datamining und statistische Prozesskontrolle zur zustandsorientierten Instandhaltung. In: VGB, Jg. 8, Nr. 10, S. 4.

Dettke, C.; Kleesatl, M.; Kuss, G.; Müller, A.; Zentgraf, R. (2009): Schnittstellenmanagement. In: Schuh, G.; Kampker, A.; Odak, R. (Hrsg.): Verfügbarkeitsorientierte Instandhaltung - Stellhebel zur Steigerung der Verfügbarkeit in produzierenden Unternehmen (Verstand). 1., Aachen: Apprimus. ISBN 978-3-940565-98-3 (Verfügbarkeit von Produktionssystemen als Dienstleistung)., S. 121–151.

Deutsches Institut für Normung (2012): DIN 31051-Grundlagen der Instandhaltung, Beuth Verlag GmbH. 2012.

Diebold, F. X. (2003): “Big Data” Dynamic Factor Models for Macroeconomic Measurement and Forecasting; A Discussion of the Papers by Lucrezia Reichlin and by Mark W. Watson. In: Dewatripont, M.; Hansen, L. P.; Turnovsky, S. J.; Econometric Society (Hrsg.): Advances in economics and econometrics: theory and applications: eighth World Congress. Cambridge: Cambridge University Press. ISBN 978-0-521-52411-7 (Econometric Society monographs)., S. 115–122.

Diekmann, A. (2001): Empirische Sozialforschung; Grundlagen, Methoden, Anwendungen. 7. durchgesichtete, Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Dippold, R.; Meier, A.; Schnider, W.; Schwinn, K. (2005): Unternehmensweites Datenmanagement: von der Datenbankadministration bis zum Informationsmanagement. 4., überarb. und erw. Aufl, Braunschweig: Vieweg. ISBN 978-3-528-35661-3.

Domschke, W.; Scholl, A. (2005): Grundlagen der Betriebswirtschaftslehre: eine Einführung aus entscheidungsorientierter Sicht. 3., verb. Aufl, Berlin: Springer. ISBN 978-3-540-25047-0.

Dorschel, J.; Dorschel, W.; Föhl, U.; van Geenen, W.; Hertweck, D.; Kinitzki, M.; Küller, P.; Lanquillon, C.; Mallow, H.; März, L.; Omri, F.; Schacht, S.; Stremler, A.; Theobald, E. (2015): Wirtschaft. In: Dorschel, J. (Hrsg.): Praxishandbuch Big Data. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN 978-3-658-07288-9, S. 15–166.

Drechsler, D. (2018): Predicitve Analytics - Eine etwas differenziertere Betrachtung eines kritischen Themas. In: Breyer-Mayländer, T. (Hrsg.): Das Streben nach Autonomie; Reflexionen zum digitaen Wandel. 1. Auflage, Baden-Baden: Nomos. , S. 237–263.

Eickelmann, M.; Wiegand, M.; Deuse, J.; Bernerstätter, R. (2019): Bewertungsmodell zur Analyse der Datenreife: Herleitung des Reifegrads der Datenqualität für die Anwendung maschineller Lernverfahren in der industriellen Produktion. In: ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb, Jg. 114, Nr. 1–2, S. 29–33.

Emran, N. A. (2015): Data Completeness Measures. In: Abraham, A.; Muda, A. K.; Choo, Y.-H. (Hrsg.): Pattern Analysis, Intelligent Security and the Internet of Things. Cham: Springer International Publishing. ISBN 978-3-319-17397-9, S. 117–130.

Ereth, J.; Kemper, H.-G. (2016): Business Analytics und Business Intelligence. In: Controlling, Jg. 28, Nr. 8–9, S. 458–464.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996): From Data Mining to Knowledge Discovery in Databases. In: AI Magazine, Jg. 17, Nr. 3, S. 37–54.

Frehe, V.; Adelmeyer, T.; Teuteberg, F. (2016): Eine Balanced Scorecard für das systematische Datenqualitätsmanagement im Kontext von Big Data. In: Nissen, V.; Stelzer, D.; Straßburger, S.; Danie, F. (Hrsg.): Multikonferenz Wirtschaftsinformatik (MKWI) 2016: Technische Universität Ilmenau 09.-11. März 2016. Ilmenau: Universitätsverlag. ISBN 978-3-86360-132-4, S. 143–154.

Gartzen, T.; Kempf, M.; Kupke, D.; Simons, H.-P. (2009): Instandhaltungsstrategie. In: Schuh, G.; Kampker, A.; Odak, R. (Hrsg.): Verfügbarkeitsorientierte Instandhaltung - Stellhebel zur Steigerung der Verfügbarkeit in produzierenden Unternehmen (Verstand). 1., Aachen: Apprimus. ISBN 978-3-940565-98-3 (Verfügbarkeit von Produktionssystemen als Dienstleistung)., S. 39–60.

Gatica, C. P.; Boschmann, A. (2019): Enabling Self-Diagnosis of Automation Devices through Industrial Analytics. In: Beyerer, J.; Kühnert, C.; Niggemann, O. (Hrsg.): Machine Learning for Cyber Physical Systems. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-662-58484-2, S. 107–115.

Geißler, P.; Nemeth, T.; Sihn, W. (2016): Smart Factory bedarf Smart Maintenance; Zu Smart Maintenance durch intelligente Instandhaltungssysteme. In: Biedermann, H. (Hrsg.): Lean smart maintenance: Konzepte, Instrumente und Anwendungen für eine effiziente und intelligente Instandhaltung. Köln: TÜV Media. ISBN 978-3-7406-0096-9 (Praxiswissen für Ingenieure - Instandhaltung)., S. 9–18.

Goeken, M. (2006): Entwicklung von Data-Warehouse-Systemen: Anforderungsmanagement, Modellierung, Implementierung. 1. Aufl, Wiesbaden: Dt. Univ.-Verl. ISBN 978-3-8350-0325-5.

Goel, S.; Bhat, J. M.; Weber, B. (2013): End-to-End Process Extraction in Process Unaware Systems. In: La Rosa, M.; Soffer, P. (Hrsg.): Business Process Management Workshops: Bpm 2012 International Workshops, Tallinn, Estonia, September 3, 2012, Revised Papers. Berlin: Springer. ISBN 978-3-642-36284-2 (Lecture Notes in Business Information Processing; 132)., S. 162–173.

Grande, M. (2011): 100 Minuten für Anforderungsmanagement: kompaktes Wissen nicht nur für Projektleiter und Entwickler. 1. Aufl, Wiesbaden: Vieweg + Teubner. ISBN 978-3-8348-1431-9.

Grosser, T. (2013): Data Governance; Daten effizienter nutzen, BARC GmbH. 2013.

Gröger, C. (2015): Advanced Manufacturing Analytics: datengetriebene Optimierung von Fertigungsprozessen. 1. Aufl, Lohmar Köln: Eul. ISBN 978-3-8441-0420-2.

Haberfellner, R.; Büchel, A.; Nagel, P.; von Massow, H.; Becker, M. (2012): Systems Engineering: Grundlagen und Anwendung. 12., aktualisierte Aufl, Zürich: Orell Füssli. ISBN 978-3-280-04068-3.

Haegemans, T.; Snoeck, M.; Lemahieu, W. (2018): Entering data correctly: An empirical evaluation of the theory of planned behaviour in the context of manual data acquisition. In: Reliability Engineering & System Safety, Jg. 178, S. 12–30.

Halford, G. S.; Baker, R.; McCredden, J. E.; Bain, J. D. (2005): How Many Variables Can Humans Process? In: Psychological Science, Jg. 16, Nr. 1, S. 70–76.

Hall, P.; Dean, J.; Kabul, I. K.; Silva, J. (2014): An overview of machine learning with SAS® enterprise miner™, SAS Institute Inc. 2014.

Handl, A. (2010): Multivariate Analysemethoden. 2. Auflage, Berlin: Springer-Verlag. ISBN 978-3-642-14986-3.

Harrach, H. (2010): Risiko-Assessments für Datenqualität: Konzept und Realisierung. 1. Aufl, Wiesbaden: Vieweg + Teubner. ISBN 978-3-8348-1344-2.

Hastings, N. A. J. (2015): Physical asset management. With an Introduction to ISO55000. 2. Aufl., New York, NY: Springer Berlin Heidelberg. ISBN 978-3-319-14776-5.

Hazen, B. T.; Boone, C. A.; Ezell, J. D.; Jones-Farmer, L. A. (2014): Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. In: International Journal of Production Economics, Jg. 154, S. 72–80.

Hedderich, J.; Sachs, L. (2016): Angewandte Statistik: Methodensammlung mit R. 15., überarbeitete und erweiterte Auflage, Berlin Heidelberg: Springer Spektrum. ISBN 978-3-662-45690-3.

Heidel, R.; Hoffmeister, M.; Hankel, M.; Döbrich, U. (2017): Industrie4.0 Basiswissen RAMI4.0: Referenzarchitekturmodell mit Industrie4.0-Komponente. 1. Auflage, Berlin Wien Zürich: Beuth Verlag GmbH. ISBN 978-3-410-26482-8.

Heil, M. (1995): Entstörung betrieblicher Abläufe. Wiesbaden: Dt. Univ.-Verl. [u.a.]. ISBN 978-3-8244-6100-4.

Heimes, H.; Kampker, A.; Bührer, U.; Krotil, S. (2019): Potenziale und Hürden von Data Analytics in der Serienfertigung. In: Industrie 4.0 Management, Jg. 2019, Nr. 1, S. 57–60.

Heinrich, B.; Klier, M. (2009): Die Messung der Datenqualität im Controlling - Ein metrikbasierter Ansatz und seine Anwendung im Kundenwertcontrolling. In: Controlling & Management Review, Jg. 53, Nr. 1, S. 34–42.

Heinrich, B.; Klier, M. (2015): Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement. In: Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. (Hrsg.): Daten- und Informationsqualität; Auf dem Weg zur Information Excellence. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN 978-3-658-09213-9, S. 49–67.

Henning, J. (2009): Wissensmanagement. In: Schuh, G.; Kampker, A.; Odak, R. (Hrsg.): Verfügbarkeitsorientierte Instandhaltung - Stellhebel zur Steigerung der Verfügbarkeit in produzierenden Unternehmen (Verstand). 1., Aachen: Apprimus. (Verfügbarkeit von Produktionssystemen als Dienstleistung)., S. 153–181.

Hettich, S.; Hippner, H. (2001): Assoziationsanalyse. In: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.-D. (Hrsg.): Handbuch Data Mining im Marketing. Knowledge Discovery in Databases. Wiesbaden: Vieweg. , S. 427–463.

Hevner, A. R. (2007): A Three Cycle View of Design Science Research. In: Scandinavian Journal of Information Systems, Jg. 19, Nr. 2, S. 87–92.

Hevner, A. R.; March, S. T.; Park, J.; Ram, S. (2004): Design Science in Information Systems Research. In: MIS Quarterly, Jg. 28, Nr. 1, S. 75–105.

Homburg, C. (1998): Quantitative Betriebswirtschaftslehre: Entscheidungsunterstützung durch Modelle ; mit Beispielen, Übungsaufgaben und Lösungen. 2., überarb. und erw. Aufl, Wiesbaden: Gabler. ISBN 978-3-409-23417-7.

Horn, G. (2012): Cause Mapping-Ereignisanalyse zum Aufdecken und Lösen von Schwachstellen. In: Der Instandhaltungs-Berater. 37. Aktualisierungsauflage, Köln: TÜV Media GmbH. , S. 31.

Huang, J.; Esbensen, K. H. (2000): Applications of Angle Measure Technique (AMT) in image analysis-Part I. A new methodology for in situ powder characterization. In: Chemometrics and Intelligent Laboratory Systemso. Jg., , Nr. 54, S. 1–19.

Jochem, R. (2010): Was kostet Qualität - Wirtschaftlichkeit von Qualität ermitteln. Müchen: Carl Hanser.

Jockisch, M.; Rosendahl, J. (2010): Klassifikation von Modellen. In: Bandow, G.; Holzmüller, H. H. (Hrsg.): ”Das ist gar kein Modell!”: unterschiedliche Modelle und Modellierungen in Betriebswirtschaftslehre und Ingenieurwissenschaften. 1. Auflage, Wiesbaden: Gabler Research. ISBN 978-3-8349-1842-0 (Gabler Research)., S. 22–52.

Jodlbauer, H.; Schagerl, M. (2016): Reifegradmodell Industrie 4.0 - Ein Vorgehensmodell zur Identifikation von Industrie 4.0 Potentialen. In: Mayr, H. C.; Pinzger, M. (Hrsg.): Informatik 2016. Bonn: Gesellschaft für Informatik e.V. (Lecture Notes in Informatics (LNI))., S. 1473–1487.

Kamsu-Foguem, B.; Rigal, F.; Mauget, F. (2013): Mining association rules for the quality improvement of the production process. In: Expert Systems with Applications, Jg. 40, Nr. 4, S. 1034–1045.

Kantardzic, M. (2011): Data mining: concepts, models, methods, and algorithms. 2nd ed, Hoboken, N.J: John Wiley : IEEE Press. ISBN 978-0-470-89045-5.

Karim, R.; Westerberg, J.; Galar, D.; Kumar, U. (2016): Maintenance Analytics – The New Know in Maintenance. In: IFAC-PapersOnLine, Jg. 49, Nr. 28, S. 214–219.

Ketteler, D.; König, C. (2017): Lean 4.0-Schlank durch Digitalisierung. Frankfurt am Main: BearingPoint GmbH. 2017.

Khazraei, K. (2011): Design, organization and implementation of a methods pool and an application systematics for condition based maintenance. Dortmund: Verl. Praxiswissen. ISBN 978-3-86975-042-2.

Kiem, R. (2016): Qualität 4.0: QM, MES und CAQ in digitalen Geschäftsprozessen der Industrie 4.0. München: Hanser. ISBN 978-3-446-44736-3.

Kinz, A.; Bernerstätter, R. (2016): Instandhaltungsoptimierung mittels Lean Smart Maintenance - Einführung des Lean Smart Maintenance Ansatzes. In: Biedermann, H. (Hrsg.): Lean Smart Maintenance: Konzepte, Instrumente und Anwendungen für eine effiziente und intelligente Instandhaltung. Köln: TÜV Media. (Praxiswissen für Ingenieure - Instandhaltung)., S. 61–100.

Kleindienst, B.; Bernerstätter, R. (2015): Kennzahlen in Smart Maintenance - Entwicklung eines Kennzahlen-Cockpits für die Instandhaltung unterstützt durch Datenanalysemethoden. In: Smart Maintenance - Intelligente, lernorientierte Instandhaltung. Köln: TÜV Media GmbH. (Praxiswissen Instandhaltung).

Kletti, J. (2015): MES - Manufacturing Execution System. Moderne Informationstechnologie unterstützt die Wertschöpfung. 2. Auflage, Berlin Heidelberg: Springer Vieweg. ISBN 978-3-662-46901-9.

Klier, M. (2008): Metriken zur Bewertung der Datenqualität – Konzeption und praktischer Nutzen. In: Informatik-Spektrum, Jg. 31, Nr. 3, S. 223–236.

Klier, M.; Heinrich, B. (2016): Datenqualität als Erfolgsfaktor im Business Analytics. In: Controlling, Jg. 28, Nr. 8–9, S. 488–494.

Kopcso, D.; Pachamanova, D. (2018): Case-Managing Staffing Inefficiencies Using Analytics (B): Business Value in Predictive and Prescriptive Analytics Models. In: INFORMS Transactions on Education, Jg. 19, Nr. 1, S. .

Kotu, V.; Deshpande, B. (2015): Predictive analytics and data mining: concepts and practice with RapidMiner. Amsterdam: Elsevier/Morgan Kaufmann, Morgan Kaufmann is an imprint of Elsevier. ISBN 978-0-12-801460-8.

Krcmar, H. (2015): Informationsmanagement. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-662-45862-4.

Krogstie, J.; Lindland, O. I.; Sindre, G. (1995): Defining quality aspects for conceptual models. In: Falkenberg, E. D.; Hesse, W.; Olivé, A. (Hrsg.): Information System Concepts. Boston, MA: Springer US. ISBN 978-1-5041-2895-7, S. 216–231.

Kuhn, M.; Johnson, K. (2013): Applied predictive modeling. New York: Springer. ISBN 978-1-4614-6848-6.

Lahrmann, G.; Stroh, F. (2008): Systemarchitektur für die Informationslogistik. In: Dinter, B.; Winter, R. (Hrsg.): Integrierte Informationslogistik. Berlin: Springer. ISBN 978-3-540-77577-5 (Business Engineering)., S. 137–165.

Lanza, G.; Nyhuis, P.; Ansari, S. M.; Kuprat, T.; Liebrecht, C. (2016): Befähigungs- und Einführungsstrategien für Industrie 4.0: Vorstellung eines reifegradbasierten Ansatzes zur Implementierung von Industrie 4.0. In: ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb, Jg. 111, Nr. 1–2, S. 76–79.

Lee, G.-Y.; Kim, M.; Quan, Y.-J.; Kim, M.-S.; Kim, T. J. Y.; Yoon, H.-S.; Min, S.; Kim, D.-H.; Mun, J.-W.; Oh, J. W.; Choi, I. G.; Kim, C.-S.; Chu, W.-S.; Yang, J.; Bhandari, B.; Lee, C.-M.; Ihn, J.-B.; Ahn, S.-H. (2018): Machine health management in smart factory: A review. In: Journal of Mechanical Science and Technology, Jg. 32, Nr. 3, S. 987–1009.

Lieber, D.; Erohin, Olga; Deuse, Jochen (2013): Wissensentdeckung im industriellen Kontext: Herausforderungen und Anwendungsbeispiele. In: Zeitschrift für wirtschaftlichen Fabrikbetrieb, Jg. 06, Nr. 108, S. 388–393.

Linders, B. (2019): CMMI V1.3 Process Areas. CMMI V1.3 Process Areas. URL: https://www.benlinders.com/tools/cmmi-v1-3-process-areas/.

Linß, G. (2018): Qualitätsmanagement für Ingenieure. 4., aktualisierte und erweiterte Auflage, München: Hanser. ISBN 978-3-446-44042-5.

Lorenz, M.; Küpper, D.; Rüßmann, M.; Heidemann, A.; Bause, A. (2016): Time to Accelerate in the Race Toward Industry 4.0, The Boston Consulting Group. URL: http://www.metalonia.com/w/documents/BCG-Time-to-Accelerate-in-the-Race-Toward-Industry-4.0-May-2016_tcm80-209674.pdf (Zugriff: 14.03.2019).

Lovell, M. C. (1983): Data Mining. In: The Review of Economics and Statistics, Jg. 65, Nr. 1, S. 1–12.

Lueth, K. L.; Patsioura, C.; Williams, Z. D.; Kermani, Z. Z. (2016): Industrial Analytics 2016/2017. The current state of data analytics usage in industrial companies. IoT Analytics GmbH. 2016.

Lusti, M. (1997): Dateien und Datenbanken: eine anwendungsorientierte Einführung ; mit 92 Tabellen. 3., vollst. überarb. und erw. Aufl, Berlin: Springer. ISBN 978-3-540-61763-1.

Madhikermi, M.; Kubler, S.; Robert, J.; Buda, A.; Främling, K. (2016): Data quality assessment of maintenance reporting procedures. In: Expert Systems with Applications, Jg. 63, S. 145–164.

Mariscal, G.; Marbán, Ó.; Fernández, C. (2010): A survey of data mining and knowledge discovery process models and methodologies. In: The Knowledge Engineering Review, Jg. 25, Nr. 02, S. 137–166.

Markl, V.; Hoeren, T.; Krcmar, H. (2013): Innovationspotentialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen (Big Data Management). 2013.

Matyas, K.; Nemeth, T.; Kovacs, K.; Glawar, R. (2017): A procedural approach for realizing prescriptive maintenance planning in manufacturing industries. In: CIRP Annals, Jg. 66, Nr. 1, S. 461–464.

Matzer, M.; Lohse, H. (2007): Dateiformate: ODF, DOCX, PSD, SMIL, WAV & Co.; Einsatz und Konvertierung. Frankfurt am Main: entwickler.press. ISBN 978-3-939084-34-1.

Mexis, N. D. (1992): Erfolgreiches Rationalisieren in der Konsumgüter-, Verpackungs- und Verfahrensindustrie. Heidelberg: Verl. für Fachliteratur Hüthig. ISBN 978-3-920993-01-0.

Mitchell, T. M. (1997): Machine learning. New York,: McGraw-Hill. ISBN 978-0-07-115467-3.

Mittag, H.-J. (2012): Statistik: Eine interaktive Einführung. 2., wesentl. überarb. Aufl, Berlin: Springer Berlin. ISBN 978-3-642-30089-9.

Moharana, U. C.; Sarmah, S. P. (2015): Determination of optimal kit for spare parts using association rule mining. In: International Journal of System Assurance Engineering and Management, Jg. 6, Nr. 3, S. 238–247.

Mosaddar, D.; Shojaie, A. A. (2013): A data mining model to identify inefficient maintenance activities. In: International Journal of System Assurance Engineering and Management, Jg. 4, Nr. 2, S. 182–192.

Munson, M. A. (2012): A study on the importance of and time spent on different modeling steps. In: ACM SIGKDD Explorations Newsletter, Jg. 13, Nr. 2, S. 65.

Nebl, T.; Prüß, H. (2006): Anlagenwirtschaft. München: Oldenbourg. ISBN 978-3-486-57961-1.

Nemeth, T.; Bernerstätter, R.; Glawar, R.; Matyas, K.; Sihn, W. (2015): Instandhaltung 4.0: Sicherstellung von Produktqualität und Anlagenverfügbarkeit durch einen echtzeitbasierten Instandhaltungsleitstand. In: ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb, Jg. 110, Nr. 9, S. 569–573.

Ninck, A.; Bürki, L.; Hungerbühler, R.; Mühlemann, H. (1998): Systemik: integrales Denken, Konzipieren und Realisieren. 2., überarb. Aufl, Zürich: Verl. Industrielle Organisation. ISBN 978-3-85743-994-0.

North, K. (2011): Wissensorientierte Unternehmensführung: Wertschöpfung durch Wissen. 5., aktualisierte und erweiterte Auflage, Wiesbaden: Gabler. ISBN 978-3-8349-2538-1.

Olson, J. E. (2003): Data quality - the accuracy dimension. San Francisco: Morgan Kaufmann. ISBN 978-1-55860-891-7.

Otto, B. (2015): Quality and Value of the Data Resource in Large Enterprises. In: Information Systems Management, Jg. 32, Nr. 3, S. 234–251.

Otto, B.; Hinderer, H. (2009): Datenqualitätsmanagement im Lieferanten-Controlling: Fallbeispiele, Architekturentwurf und Handlungsempfehlungen. In: Controlling & Management, Jg. 53, Nr. 1, S. 21–29.

Otto, B.; Österle, H. (2016): Corporate Data Quality: Voraussetzung erfolgreicher Geschäftsmodelle. Berlin Heidelberg: Springer Gabler. ISBN 978-3-662-46805-0.

Pascual, D. G. (2015): Artificial intelligence tools: decision support systems in condition monitoring and diagnosis. Boca Raton, Florida: CRC Press. ISBN 978-1-4665-8405-1.

Paulk, M. C.; Curtis, B.; Chrissis, M. B.; Weber, C. V. (1995): Capability Maturity Model for Software - Version 1.1. Pittsburgh: Carnegie Mellon University. 1995.

Pawellek, G. (2016): Integrierte Instandhaltung und Ersatzteillogistik: Vorgehensweisen, Methoden, Tools. 2. Auflage, Berlin Heidelberg: Springer Vieweg. ISBN 978-3-662-48666-5.

Petersohn, H. (2005): Data Mining: Verfahren, Prozesse, Anwendungsarchitektur. München: Oldenbourg. ISBN 978-3-486-57715-0.

Picot, A.; Reichwald, R. (1991): Informationswirtschaft. In: Industriebetriebslehre; Entscheidungen im Industriebetrieb. 9., vollstandig neu bearbeitete und erweiterte Auflage, Wiesbaden: Gabler. , S. 241–393.

Promotorengruppe Kommunikation der Forschungsunion Wirtschaft – Wissenschaft (2013): Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0 Deutschlands Zukunft als Produktionsstandort sichern ; Abschlussbericht des Arbeitskreises Industrie 4.0. Frankfurt am Main: Geschäftsstelle der Plattform Industrie 4.0. 2013.

Reder, L.; Steven, M.; Klünder, T. (2018): Industrie 4.0-Readiness von Supply Chain-Netzwerken; Qualitative und quantitative Analyse am Beispiel der Automobilindustrie. In: Industrie 4.0 Management, Jg. 2018, Nr. 5, S. 11–16.

Redman, T. C. (1996): Data quality for the information age. Boston: Artech House. ISBN 978-0-89006-883-0.

Rohweder, J. P.; Kasten, G.; Malzahn, D.; Piro, A.; Schmid, J. (2015): Informationsqualität – Definitionen, Dimensionen und Begriffe. In: Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. (Hrsg.): Daten- und Informationsqualität. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN 978-3-658-09213-9, S. 25–46.

Runkler, T. A. (2010): Data Mining: Methoden und Algorithmen intelligenter Datenanalyse ; Mit 72 Abbildungen und 7 Tabellen. 1. Aufl, Wiesbaden: Vieweg+Teubner. ISBN 978-3-8348-0858-5.

Ryll, F.; Freund, C. (2010): Grundlagen der Instandhaltung. In: Schenk, M. (Hrsg.): Instandhaltung technischer Systeme: Methoden und Werkzeuge zur Gewährleistung eines sicheren und wirtschaftlichen Anlagenbetriebs. Berlin: Springer. ISBN 978-3-642-03948-5, S. 23–101.

Sappelli, M.; de Boer, M. H. T.; Smit, S. K.; Bomhof, F. (2017): A Vision on Prescriptive Analytics. In: Gudivada, V. N.; Bhulai, S.; di Buono, M. P. (Hrsg.): ALLDATA 2017;The Third International Conference on Big Data, Small Data, Linked Data and Open Data. Venedig: IARA. , S. 45–50.

Scheller, T. (2017): Auf dem Weg zur agilen Organisation: wie Sie Ihr Unternehmen dynamischer, flexibler und leistungsfähiger gestalten. München: Verlag Franz Vahlen. ISBN 978-3-8006-5271-6.

Schenkendorf, R.; Böhm, T. (2015): Aspekte einer datengetriebenen, zustandsabhängigen Instandhaltung - (Teil 2) Vom Merkmal zur Fehlerdetektion. In: EI-Eisenbahningenieuro. Jg., , Nr. März, S. 21–25.

Schmidt, R.; Mohring, M. (2013): Strategic Alignment of Cloud-Based Architectures for Big Data. Proceedings of 2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops (EDOCW), Vancouver, BC, Canada, 2013. Vancouver, BC, Canada: IEEE. ISBN 978-1-4799-3048-7.

Schmitz, S.; Krenge, J. (2014): Big Data im Sörungsmanagement. Plödereder, E.; Grunske, L.; Schneider, E.; Ull, D. (Hrsg.), Bonn, 2014. Bonn: Gesellschaft für Informatik e.V.

Schuh, G.; Anderl, R.; ten Hompel, M.; Wahlster, W. (2017): Industrie 4.0 Maturity Index. Die digitale Transformation von Unternehmen gestalten (acatech STUDIE). Herbert Utz Verlag.

Schulmeyer, C. (2015): Big Data-Analse auf Basis technischer Methoden und Systeme. In: Dorschel, J. (Hrsg.): Praxishandbuch Big Data: Wirtschaft - Recht - Technik. Wiesbaden: Springer Gabler. ISBN 978-3-658-07288-9, S. 307–329.

Schumacher, A.; Erol, S.; Sihn, W. (2016): A Maturity Model for Assessing Industry 4.0 Readiness and Maturity of Manufacturing Enterprises. In: Procedia CIRP, Jg. 52, S. 161–166.

Schweitzer, M. (1994): Gegenstand der Industrieberiebslehre. In: Schweitzer, M. (Hrsg.): Industriebetriebslehre: das Wirtschaften in Industrieunternehmungen. 2., völlig überarb. und erw. Aufl, München: Vahlen. ISBN 978-3-8006-1755-5 (Vahlens Handbücher der Wirtschafts- und Sozialwissenschaften)., S. 3–60.

Schweitzer, M.; Küpper, H.-U. (1997): Produktions- und Kostentheorie: Grundlagen - Anwendungen. 2., vollst. überarb. und wesentlich erw. Aufl, Wiesbaden: Gabler. ISBN 978-3-409-12167-5.

Schütte, R. (1998): Grundsätze ordnungsmäßiger Referenzmodellierung: Konstruktion konfigurations- und anpassungsorientierter Modelle. Wiesbaden: Gabler. ISBN 978-3-409-12843-8.

Siebert, H. G. (2010): Prozessverbesserung in der IT mit SPICE. In: OBJEKTspektrumo. Jg., , Nr. 01, S. 76–81.

Sonka, M.; Hlavac, V.; Boyle, R. (2015): Image processing, analysis, and machine vision. Fourth edition, Stamford, CT, USA: Cengage Learning. ISBN 978-1-133-59360-7.

Staar, B.; Kück, M.; Ait Alla, A.; Lütjen, M.; Freitag, M.; Simic, A. (2017): Statistische Detektion von Anomalien in Bilddaten von Mikrobauteilen - Statistische Defekterkennung mittels Hauptkomponentenanalyse. In: Industrie 4.0 Management, Jg. 33, Nr. 2, S. 52–55.

Stachowiak, H. (1973): Allgemeine Modelltheorie. Wien: Springer.

Stachowiak, H. (1980): Der Modellbegriff in der Erkenntnistheorie. In: Zeitschrift für allgemeine Wissenschaftstheorie, Jg. 11, Nr. 1, S. 53–68.

Stachowiak, H. (1992): Modell. In: Seiffert, H.; Radnitzky, G. (Hrsg.): Handlexikon der Wissenschaftstheorie. 2. Aufl., 7.-9. Tsd, München: Dt. Taschenbuch-Verl. ISBN 3-423-04586-5 (dtv dtv Wissenschaft; 4586)., S. 219–222.

Stodder, D. (2016): Improving Data Preparation for Business Analytics; Applying Technologies and Methods for Establishing Trusted Data Assets for More Productive Users. Renton: tdwi. 2016.

Strunz, M. (2012): Instandhaltung: Grundlagen - Strategien - Werkstätten. Berlin: Springer. ISBN 978-3-642-27389-6.

Taheri, M.; Mahdavi, A. (2018): Structured representation of monitored occupancy data. In: Bauphysik, Jg. 40, Nr. 6, S. 434–440.

Terrizzano, I.; Schwarz, P.; Roth, M.; Colino, J. E. (2015): Data Wrangling: The Challenging Yourney from the Wild to the Lake. Proceedings of 7th Biennial Conference on Innovative Data Systems Research, Kalifornien, 2015. Kalifornien: 9.

Troitzsch, K. G. (1990): Modellbildung und Simulation in den Sozialwissenschaften. Opladen: Westdeutscher Verlag. ISBN 3-531-12150-2.

Töllner, A.; Jungmann, T.; Bücker, M.; Brutscheck, T. (2010): Modelle und Modellierung - Terminologie, Funktionen und Nutzen. In: Bandow, G.; Holzmüller, H. H. (Hrsg.): ”Das ist gar kein Modell!”: unterschiedliche Modelle und Modellierungen in Betriebswirtschaftslehre und Ingenieurwissenschaften. 1. Auflage, Wiesbaden: Gabler Research. ISBN 978-3-8349-1842-0 (Gabler Research)., S. 3–21.

Vaishnavi, V.; Kuechler, W. (2015): Design science research methods and patterns: innovating information and communication technology. Second edition, Boca Raton: CRC Press, Taylor & Francis Group. ISBN 978-1-4987-1525-6.

Van der Aalst, W. (2016): Process mining: data science in action. 2nd edition, New York, NY: Springer Berlin Heidelberg. ISBN 978-3-662-49850-7.

VDI Verlag (2019): Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution. vdi-nachrichten.com. URL: https://www.vdi-nachrichten.com/Technik-Gesellschaft/Industrie-40-Mit-Internet-Dinge-Weg-4-industriellen-Revolution (Zugriff: 14.03.2019).

VDMA; Fraunhofer IOSB-INA (2017): Industrie 4.0 Kommunikation mit OPC-UA; Leitfaden zur Einführung in den Mittelstand, VDMA Verlag. URL: https://industrie40.vdma.org/documents/4214230/16617345/1492669959563_2017_Leitfaden_OPC_UA_LR.pdf/f4ddb36f-72b5-43fc-953a-ca24d2f50840 (Zugriff: 03.02.2019).

Vom Brocke, J.; Grob, H. L. (2015): Referenzmodellierung: Gestaltung und Verteilung von Konstruktionsprozessen. 2., unveränderte Auflage, Berlin: Logos Verlag. ISBN 978-3-8325-0179-2.

Voß, S.; Gutenschwager, K. (2001): Informationsmanagement: mit 25 Tabellen. Berlin: Springer. ISBN 978-3-540-67807-6.

Wang, R. Y. (1998): A product perspective on total data quality management. In: Communications of the ACM, Jg. 41, Nr. 2, S. 58–65.

Wang, R. Y.; Strong, D. M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers. In: Journal of Management Information Systems, Jg. 12, Nr. 4, S. 5–33.

Warnecke, H.-J. (1992): Bedeutung der Funktion ”Instandhaltung”. In: Warnecke, H.-J. (Hrsg.): Handbuch Instandhaltung, Band 1. Instandhaltungsmanagement. 2., völlig überarb. Aufl, Köln: Verl. TÜV Rheinland. ISBN 978-3-88585-822-5 (Handbuch Instandhaltung; 1)., S. 3–16.

Weigel, N. (2015): Datenqualitätsmanagement - Steigerung der Datenqualität mit Methode. In: Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. (Hrsg.): Daten- und Informationsqualität: Auf dem Weg zur Information Excellence. 3., erweiterte Auflage, Wiesbaden: Springer Vieweg. ISBN 978-3-658-09213-9, S. 69–86.

Werner, M. (2004): Einflussfaktoren des Wissenstransfers in wissensintensiven Dienstleistungsunternehmen. Wiesbaden: Deutscher Universitätsverlag. ISBN 978-3-8244-8244-3.

Wirth, R.; Hipp, J. (2000): CRISP-DM: Towards a standard process model for data mining. Proceedings of Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, Manchester, 2000. Manchester: Blackpool.

Witten, I. H.; Frank, E. (2005): Data mining: practical machine learning tools and techniques. 2nd ed, Amsterdam ; Boston, MA: Morgan Kaufman. ISBN 978-0-12-088407-0.

Witten, I. H.; Frank, E.; Hall, M. A. (2011): Data mining: practical machine learning tools and techniques. 3rd ed, Burlington, MA: Morgan Kaufmann. ISBN 978-0-12-374856-0.

Woodall, P.; Gao, J.; Parlikad, A.; Koronios, A. (2015): Classifying Data Quality Problems in Asset Management. In: Tse, P. W.; Mathew, J.; Wong, K.; Lam, R.; Ko, C. N. (Hrsg.): Engineering Asset Management - Systems, Professional Practices and Certification. Processdings of 8th World Congress on Engineering Asset Management (WCEAM 2013) & the 3re International Conference on Utility Management & Syfety (ICUMAS). Cham: Springer International Publishing. ISBN 978-3-319-09506-6, S. 321–334.

Wöhe, G.; Döring, U.; Brösel, G. (2016): Einführung in die allgemeine Betriebswirtschaftslehre. 26., überarbeitete und aktualisierte Auflage, München: Verlag Franz Vahlen. ISBN 978-3-8006-5000-2.

Yan, J. (2015): Machinery prognostics and prognosis oriented maintenance management. Singapore: John Wiley & Sons Singapore Pte. Ltd. ISBN 978-1-118-63872-9.

ÖNORM EN ISO 9000:2015-11-15 (2015): Qualitätsmanagementsysteme – Grundlagen und Begriffe, Österreichisches Norminstitut. 2015.

Anhang

A. Fragenkatalog

Generelle Fragen zur Einordnung der Ziele und des Betriebs

1. Was ist das Ziel des Projekts?
2. Welche Motivation gibt es für das Projekt?
3. Welche Daten werden im Projektbereich erhoben?
    a. Produktionsdaten
    b. Instandhaltungsdaten
    c. Qualitätsdaten
    d. Maschinendaten
        i. SPS
        ii. Sensormessungen
4. Wo werden die Daten erhoben?
    a. Direkt am Ort des Geschehens
    b. In einem Messraum
    c. Nicht direkt vor Ort
5. Es existiert eine Data Governance?
    a. Es gibt Rollen für Datenverantwortliche usw.?
    b. Es gibt einen Prozess die Datenqualität auch von Bewegungsdaten hoch zu halten?

Fragen zur Datenerfassung

6. Die Datenerfassung wird strukturiert vorgenommen. Es wurde festgehalten:
    a. Wo?
    b. Was?
    c. Wie?
    d. Wie oft?

    gemessen wird.
7. Wie werden die Daten erfasst?
    a. Automatisch durch:
        i. BDE
        ii. MDE
        iii. MES
        iv. Andere
    b. Manuel durch:
        i. Papieraufzeichnungen
        ii. Mobile Devices (Messgeräte sowie Aufzeichnungsgeräte)
        iii. Terminaleingaben
        iv. Andere
8. Automatische Übertragung bei Rundgängen durch RFID oder NFC o. ä.?
    a. Ja (welche Art?)
    b. Nein
9. Wer erfasst die Daten?
    a. Mitarbeiter des Prozessschrittes (Produktionsmitarbeiter)

   b. Mitarbeiter von Supportprozessen (Instandhaltungsmitarbeiter)
   c. Dienstleister
   d. Andere
10. Wie oft werden die Daten erfasst?
   a. Bei regelmäßigen Rundgängen.
   b. Sporadisch
   c. Werden mit Abtastrate X erfasst.
   d. Bei gewisser Wertänderung um X in Zeit Y
11. Gibt es bekannte Probleme bei der Datenerfassung?
   a. Ja
   b. Nein, Probleme werden erkannt durch:
      i. Organisatorische Maßnahmen (Kontrolle durch den Vorgesetzten)
      ii. Automatische Erkennungsverfahren (Logische Regeln zu den Messwerten und den Zusammenhängen)
12. Vertrauen Sie den Daten?
   a. Nein, gar nicht
   b. Teilweise
   c. Großteils
   d. Ja, durchwegs

Fragen zur Datenbereitstellung (Datenspeicherung und –transfer)

13. Wie erfolgt ein Datenaustausch?
   a. Echtzeitübertragung
   b. Bulkübertragung
      i. Aus einem mobile Device
      ii. Aus einem Terminal
      iii. Andere
   c. Papiernachschreibung
14. Gibt es Standardschnittstellen zur Übertragung
   a. Ja,
      i. OPC/UA
      ii. Andere
   b. Nein, das System ist proprietär und nicht mit anderen kompatibel
      i. Export ist möglich
      ii. Export ist nicht möglich
15. Existiert ein Referenzarchitekturmodell?
   a. Ja, welches
      i. RAMI 4.0
      ii. Andere
   b. Nein
16. Die Daten sind entlang der Wertschöpfung einheitlich und durchgehend erfasst?
   a. Ja
   b. Nein
17. Die Daten werden entlang der Hierarchie aggregiert (Shopfloor bis Planung) bzw. dekompositioniert (Von der Planung heruntergebrochen auf die Feinebene)?
   a. Ja
   b. Nein
18. Wo werden die Daten gespeichert?
   a. Papierordner

   b. Direkt an der Anlage (z. B. Logfiles)
   c. Elektronisch auf einem Server (in einem Ordner)
   d. Lokal auf einem Rechner
   e. In der Datenbank (SQL, Oracle, Access, ERP)
   f. In einem Data-Warehouse mit/ohne Data-Marts und/oder OLAP Funktionalitäten
   g. In einem Data Lake
   h. In der Cloud
   i. Keine Speicherung (Daten werden nur angezeigt und dann „gelöscht“)
   j. Andere
19. Ist einer der Speicherorte ein Single Point of Truth?
   a. Ja
   b. Nein
20. Wechseln die Datenquellen im Lebenszyklus der Daten?
   a. Beispiel: Papieraufschreibung – Übertragung in Exceldatei – Übertragung in SAP
   b. Mobiles Geräte – spätere Synchronisation
   c. Andere Abläufe
21. Werden die Daten an der Anlage bei der Erfassung bereits vorverarbeitet?
   a. Nein
   b. Ja
      i. Fotos/Videos/Audio werden strukturierte Features extrahiert
      ii. Feature Extraktion anderer Art
22. Sonstige Probleme bei der Speicherung, Übertragung oder dem Zugang der Daten

Fragen zu den Datenformaten

23. In welchen Formaten liegen die Daten vor?
   a. Excel
   b. CSV
   c. TXT
   d. Word
   e. PDF
   f. XML und Derivate
   g. HTML und Derivate
   h. Big Data Formate [HDF(S)]
   i. Bildformate
   j. Audioformate
   k. Videoformate
   l. Andere Formate z. B. proprietäre Formate
   m. Keine Formate – Datenübertragung direkt über Schnittstellen
24. Gibt es Datenformate die nur mit einer bestimmten Software geöffnet werden können? [Hinweis auf ein proprietäres Format]
   a. Ja
   b. Nein

Fragen zur Datencodierung und -darstellung

25. Werden Fließtexte eingegeben?
   a. Nein
   b. Ja

i. Beliebiger Freitext
ii. Texte/Wörter aus einer Lister auswählen
iii. Texte/Wörter werden nicht ausgewählt, sondern die Standards werden eingetippt

26. Existiert ein Codierungssystem?
    a. Nein
    b. Ja, die Codes sind:
        i. Numerisch
        ii. Nicht-Numerisch
        iii. Alphanumerisch
        iv. Das Code-System ist in sich hierarchisch aufgebaut
        v. Die Codes sind atomar abgelegt
27. Gibt eine Ausschluss- oder Vorschlagslogik zwischen Rückmeldeort oder Rückmeldeursache und den Codes?
    a. Ja
    b. Nein
28. Ist das Codesystem bei allen Quellen durchgängig?
    a. Ja
    b. Nein – Es ist eine alle umfassende Übersetzungstabelle vorhanden
        i. Ja
        ii. Nein
    a. Wenn nein, gibt es eine alle umfassende Übersetzungstabelle?
29. Welches Skalenniveau haben die Daten?
    a. Metrisch (Physikalische Werte)
    b. Nominal (Schadenscodes, Rezepturen, Rüstprogramme, …)
    c. Ordinal (Gibt es in den Einträgen eine Hierarchie)
    d. Binär (Es wird ja/nein, o. Ä. eingetragen)
30. Werden Metadaten abgebildet?
    a. Ja
    b. Nein
31. Es werden Event-IDs erfasst, die Aufträge, Produkte, Chargen und ähnliches über die Zeit der Behandlung identifizieren?
    a. Ja
    b. Nein

Fragen zum Datenumfang

32. Wie lange werden die Daten gespeichert?
    a. 3 Monate
    b. 6 Monate
    c. 9 Monate
    d. 1 Jahr
    e. Kürzer
    f. Länger
33. Seit wann werden die Daten gespeichert?
34. Wieviele Attribute werden aufgezeichnet?
    a. 3 und weniger
    b. Bis 5
    c. mehr
35. Sind die Rohdaten erfasst?
    a. ja

   b. nein die Daten sind aggregiert
      i. Aggregation erfolgt sofort
      ii. Aggregation erfolgt nach 1 Jahr
      iii. Aggregation erfolgt früher als 1 Jahr
      iv. Aggregation erfolgt später als 1 Jahr
36. Wie oft kommt es zu Auffälligkeiten in einem Quartal?
   a. Störungen
   b. Qualitätsfehlern
   c. Produktionsverzug
   d. Anderes

Fragen zur Datenkonsistenz

37. Haben die Daten einen Zeitstempel?
   a. Nein
   b. Ja. Welcher Zeitpunkt wird erfasst? Z. B.
      i. Probenentnahme
      ii. Messung der Probe
      iii. Anlagenausfall
      iv. Wiederinbetriebnahme
      v. Meldung
      vi. Quittierung
      vii. Usw.
38. Was ist die Zeitspanne zwischen Ereignis und Aufzeichnung im System?
   a. Echtzeit
   b. Schichtende
   c. Andere
39. Werden andere zeitliche Komponenten erfasst?
   a. Dauer von Prozessschritten
   b. Dauer eines Stillstandes
   c. Andere
40. Gibt es einen Offset in der Zeiterfassung?
   a. Nein
   b. Ja und
      i. Dieser ist konstant
      ii. Dieser ist konstant in einem gewissen Rahmen
      iii. Führt nicht zu einer Umreihung von Events
      iv. Ist unbekannt und willkürlich

B. Fallbeispiel Automobilhersteller

Tabelle 30: Beitrag Automobilhersteller Datenmanagement[548]

Daten-erfassung	Betroffen	Datenbereit-stellung	Betroffen	Daten-formate	Betroffen
Manuell	Alle Daten	Papier	IH-Daten	Xlsx und csv	IH und Qualitäts-daten
Unregel-mäßig	IH-Daten	ERP System	IH-Daten	PDF	Qualitäts-daten
Nicht Digital	IH-Daten	Proprietäres System	Qualitäts-daten	Proprietäres Format	Qualitäts-daten
Regelmäßig	Qualitäts-daten	Lokaler Rechner	IH und Qualitäts-daten		
Digital	Qualitäts-daten				

Tabelle 31: Beitrag Automobilhersteller Datenstruktur[549]

Daten-darstellung	Betroffen	Daten-umfang	Betroffen	Daten Konsistenz	Betroffen
Nominale Skalierung	IH-Daten	Mehrere Jahre	IH- und Qualitäts-daten	Unterschied zwischen Datenanfall und Datenaufzeich-nung	IH-Daten
Metrische Skalierung	Qualitätsdaten	Geringe Stichp-robe relevanter Merkmale	IH-Daten	Unterschied zwischen Datenaufzeich-nung und relevantem Prozesszeit-punkt	Qualitäts-daten
Unstrukturiert	IH-Daten			Kein definierter Offset	IH- und Qualitäts-daten
Codeschema	IH-Daten aufbereitet				
Bilddatei	Stromaufnahme				
	IH-Daten				

[548] Quelle: Eigene Darstellung
[549] Quelle: Eigene Darstellung

C. Fallbeispiel Stahlverarbeiter

Tabelle 32: Beitrag Stahlverarbeiter Datenmanagement[550]

Daten-erfassung	Betroffen	Datenbereit-stellung	Betroffen	Daten-formate	Betroffen
Automatisch	Sensordaten	ERP-System	IH-Tätigkeiten	Excel	IH-Tätigkeiten
Regelmäßig	Sensordaten und IH-tätigkeiten (Wartungs-plan)	Lokales proprietäres Subsystem	Sensordaten	Proprietäres Format	Sensordaten
Manuell	IH-Tätigkeiten				
Digital	Sensordaten und IH-Tätigkeiten				

Tabelle 33: Beitrag Stahlverarbeiter Datenstruktur[551]

Daten-darstellung	Betroffen	Daten-umfang	Betroffen	Daten Konsistenz	Betroffen
Unstrukturierter Text	IH-Tätigkeiten	Aufzeichnungen über mehrere Monate (theoretisch Jahre)	Sensoraufzeichnungen	Zeitstempel auf ms genau	Sensordaten
Metrisches Skalenniveau	Sensordaten	Daten werden nach drei Monaten gelöscht. Jedoch ausreichend große Stichprobe	IH-Tätigkeiten	Zeitstempel auf Sekunden genau	IH-Tätigkeiten

[550] Quelle: Eigene Darstellung
[551] Quelle: Eigene Darstellung