

VEREIN  
DEUTSCHER  
INGENIEURE

VERBAND DER  
ELEKTROTECHNIK  
ELEKTRONIK  
INFORMATIONSTECHNIK

Implementierung und Betrieb von  
Big-Data-Anwendungen  
in der produzierenden Industrie

Datenqualität

Implementation and operation of big data  
applications in the manufacturing industry

Data quality

VDI/VDE 3714

Blatt 2 / Part 2

Ausgabe deutsch/englisch  
Issue German/English

*Die deutsche Version dieser Richtlinie ist verbindlich.*

*The German version of this standard shall be taken as authoritative. No guarantee can be given with respect to the English translation.*

Inhalt	Seite	Contents	Page
Vorbemerkung .....	2	Preliminary note.....	2
Einleitung.....	2	Introduction.....	2
<b>1 Anwendungsbereich.....</b>	<b>4</b>	<b>1 Scope.....</b>	<b>4</b>
<b>2 Normative Verweise.....</b>	<b>5</b>	<b>2 Normative references.....</b>	<b>5</b>
<b>3 Begriffe.....</b>	<b>5</b>	<b>3 Terms and definitions.....</b>	<b>5</b>
<b>4 Datenqualität.....</b>	<b>5</b>	<b>4 Data quality.....</b>	<b>5</b>
4.1 Struktur zur Beschreibung der Datenqualität.....	5	4.1 Structure for the description of data quality.....	5
4.2 Dimensionen der Datenqualität.....	6	4.2 Dimensions of data quality.....	6
4.3 Quantitative Bewertung der Datenqualität (Metrik).....	18	4.3 Quantitative evaluation of data quality (metrics).....	18
Schrifttum .....	19	Bibliography.....	19

VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA)

Fachbereich Digitale Transformation

**VDI-Handbuch Informationstechnik, Band 1: Angewandte Informationstechnik**  
**VDI/VDE-Handbuch Automatisierungstechnik**

## Vorbemerkung

Der Inhalt dieser Richtlinie ist entstanden unter Beachtung der Vorgaben und Empfehlungen der Richtlinie VDI 1000.

Alle Rechte, insbesondere die des Nachdrucks, der Fotokopie, der elektronischen Verwendung und der Übersetzung, jeweils auszugsweise oder vollständig, sind vorbehalten.

Die Nutzung dieser Richtlinie ist unter Wahrung des Urheberrechts und unter Beachtung der Lizenzbedingungen ([www.vdi.de/richtlinien](http://www.vdi.de/richtlinien)), die in den VDI-Merkblättern geregelt sind, möglich.

Allen, die ehrenamtlich an der Erarbeitung dieser Richtlinie mitgewirkt haben, sei gedankt.

Eine Liste der aktuell verfügbaren und in Bearbeitung befindlichen Blätter dieser Richtlinienreihe sowie gegebenenfalls zusätzliche Informationen sind im Internet abrufbar unter [www.vdi.de/3714](http://www.vdi.de/3714).

## Einleitung

Der Fachausschuss „Big Data“ der VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik hat sich mit der Erstellung dieser Richtlinie der Aufgabe angenommen, den ökonomischen und ökologischen Nutzen von Big Data aufzuzeigen, den Wissenstransfer über verschiedene Industrien und Branchen hinweg zu verbessern und die Implementierung und den Betrieb von Big-Data-Anwendungen in der produzierenden Industrie voranzutreiben und zu vereinheitlichen.

Diese Richtlinienreihe soll eine Orientierung über erforderliche Maßnahmen zur Big-Data-Analyse geben und aufzeigen, welche Methoden für eine zielführende Arbeit geeignet sind und welche Einschränkungen und Hindernisse bestehen. Praktikern und Praktikerinnen sollen Hinweise gegeben werden, welche Methoden und Betrachtungen für den Erfolg eines Big-Data-Projekts hinsichtlich des Einsatzes und des nachhaltigen Betriebs notwendig sind.

Die Richtlinienreihe VDI/VDE 3714 umfasst die Blätter:

Blatt 1 Durchführung von Big-Data-Projekten

**Blatt 2** Datenqualität

Blatt 3 Datenbewirtschaftung

Blatt 4 Analyseverfahrensklassen

Blatt 5 Modellierungsverfahren

Blatt 6 Validierung von Modellen

Blatt 7 Online-Anwendung von datengetriebenen Modellen

## Preliminary note

The content of this standard has been developed in strict accordance with the requirements and recommendations of the standard VDI 1000.

All rights are reserved, including those of reprinting, reproduction (photocopying, micro copying), storage in data processing systems and translation, either of the full text or of extracts.

The use of this standard without infringement of copyright is permitted subject to the licensing conditions ([www.vdi.de/richtlinien](http://www.vdi.de/richtlinien)) specified in the VDI Notices.

We wish to express our gratitude to all honorary contributors to this standard.

A catalogue of all available parts of this series of standards and those in preparation as well as further information, if applicable, can be accessed on the internet at [www.vdi.de/3714](http://www.vdi.de/3714).

## Introduction

The “Big Data” Technical Committee of the VDI/VDE Society Measurement and Automatic Control has taken on the task of drawing up this standard to demonstrate the economic and ecological benefits of big data, to improve the transfer of knowledge across different industries and sectors, and to promote and standardize the implementation and operation of big data applications in the manufacturing industry.

This series of standards is intended to provide orientation on required measures for big data analysis and to show which methods are suitable for target-oriented work and which limitations and obstacles exist. Practitioners should be given advice on which methods and considerations are necessary for the success of a big data project regarding its use and sustainable operation.

The series of standards VDI/VDE 3714 comprises the parts:

Part 1 Implementation of Big Data projects

**Part 2** Data quality

Part 3 Data management

Part 4 Analysis process classes

Part 5 Modelling procedures

Part 6 Validation of models

Part 7 Online application of data-driven models

Das vorliegende Blatt 2 betrachtet die Qualitätsdimensionen technischer Daten im Hinblick auf Generierung, Transfer, Speicherung und Nutzung in Big-Data-Anwendungen.

Die Richtlinienreihe VDI/VDE 3714 ist im Fachausschuss 7.24 „Big Data“ des Fachbereich 7 „Digitale Transformation“ der VDI/VDE-Gesellschaft für Mess- und Automatisierungstechnik (GMA) entstanden. Damit stellen die Produktion sowie die Mess- und Automatisierungstechnik die Schwerpunkte dar. In den Produktionsprozessen werden beispielsweise für Steuerungs- und Regelungsaufgaben oder für die Qualitätssicherung große Datenmengen erhoben, die mittels Datenanalyse für weitere Prozess- und Geschäftsverbesserungen genutzt werden können. Diese Richtlinienreihe gibt eine generelle Orientierung sowie Hinweise auf potenzielle Schwierigkeiten und Hürden bei der Durchführung von Big-Data-Anwendungen von der Entwicklung über die Inbetriebnahme bis zum nachhaltigen Betrieb.

Benachbart zu Big Data finden sich Themen wie das Internet der Dinge (IoT), die Vernetzung von Geräten (Smart Devices) oder die zunehmende „Rechnerallgegenwart“ (Ubiquitous Computing) sowie Begriffe, wie Business Intelligence, Data Analytics, Advanced Analytics, Data Mining, Smart Data und Data-Warehouse-Systeme, die generell die Nutzung von Daten adressieren.

Die Richtlinienreihe geht von einer generellen Verfügbarkeit aller benötigten Daten aus. Bezüglich der Datenmenge, ihrer Struktur und Integrität wird keine Annahme getroffen. Zur Diskussion und Charakterisierung der Daten helfen die sogenannten „fünf Vs“, die die einzelnen Dimensionen von Big Data bezeichnen. Die Daten werden charakterisiert durch Umfang (*Volume*), Unterschiedlichkeit (*Variety*) und ihre zeitliche Taktung (*Velocity*). Insbesondere bei industriellen Anwendungen sind die Qualität der Daten (*Validity*) und der unternehmerische Mehrwert (*Value*) relevant.

In Big-Data-Anwendungen werden sehr viele Daten aus den unterschiedlichsten Quellen vernetzt, um daraus Informationen und dann letztendlich Prozesswissen zu generieren. Damit stellen die Daten die Basis der wertschöpfenden Modellbildung in Big-Data-Projekten dar. In Bild 1 ist der Prozess der Datennutzung zur Erhöhung der Wettbewerbsfähigkeit in Form einer „Wissenstreppe“ schematisch in Anlehnung an *North* [1] dargestellt.

In diesem Zusammenhang wird deutlich, dass bei Big-Data-Projekten die Erhöhung der Wettbewerbsfähigkeit und damit der Wertschöpfung immer von den Daten abhängt. Also hat die Qualität der Daten

The following Part 2 considers the quality dimensions of technical data with respect to generation, transfer, storage, and use in big data applications.

The series of standards VDI/VDE 3714 is published in the Technical Committee 7.24 “Big Data” of the Technical Division 7 “Digital Transformation” of the VDI/VDE Society Measurement and Automatic Control (GMA). Thus, production as well as measurement and automation technology represent the focal points. In production processes, for example, large amounts of data are collected for control and regulation tasks or for quality assurance, which can be used for further process and business improvements by means of data analysis. This series of standards provides a general orientation as well as indications of potential difficulties and hurdles in the implementation of big data applications, from development through commissioning to sustainable operation.

Adjacent to big data are topics such as the Internet of things (IoT), the networking of devices (smart devices), or the increasing “computer omnipresence” (ubiquitous computing), as well as terms such as business intelligence, data analytics, advanced analytics, data mining, smart data, and data warehouse systems that generally address the use of data.

The series of standards assumes a general availability of all required data. No assumption is made regarding the amount of data, its structure and integrity. For the discussion and characterization of data, the so-called “five Vs”, which denote the individual dimensions of big data, are helpful. The data is characterized by *volume*, *variety*, and *velocity*. The quality of the data (*validity*) and the added business value (*value*) are particularly relevant for industrial applications.

In big data applications, a great deal of data from a wide variety of sources is networked to generate information and then ultimately process knowledge. The data thus forms the basis for value-creating model building in big data projects. In Figure 1, the process of using data to increase competitiveness is shown schematically in the form of a “knowledge staircase” based on *North* [1].

In this context, it becomes clear that in big data projects, increasing competitiveness and thus value creation always depends on the data. Thus, the quality of the data and the information generated from it has

und der daraus generierten Informationen einen erheblichen Einfluss auf die Güte (oder Quantität) der Wertschöpfung von Datenprojekten. Da datengetriebene Modelle mit Trainingsdaten formuliert und mit Validierungsdaten überprüft werden, hängt die Güte des Modells wesentlich von der Qualität der Trainings- und Validierungsdaten ab. Dabei beschreibt der Begriff „Datenqualität“ (siehe auch Abschnitt 4) die Ausprägung von Eigenschaften eines Datenbestands, die dessen Eignung ausmachen, festgelegte und vorausgesetzte Anforderungen zu erfüllen. Daher kann die Datenqualität nur im Zusammenhang mit der Datennutzung und der Daten-erzeugung beurteilt werden, die in VDI/VDE 3714 Blatt 3 ausführlich behandelt werden.



Bild 1. Wissensstreppe nach North

## 1 Anwendungsbereich

Mit dem Begriff „Big Data“ werden – obwohl er bereits seit einigen Jahren verwendet wird – unverändert sehr unterschiedliche Themen und Aspekte assoziiert und entsprechend in der gesellschaftlichen Diskussion differenziert diskutiert. Die immer weiter vorschreitende digitale Kommunikation, der in der Umsetzung befindliche Breitbandausbau und die überall mögliche Verarbeitungsmöglichkeit von Daten beflügeln diese Diskussion sowohl in der Öffentlichkeit als auch in der Fachwelt. Die Themen reichen von Datenschutz über Datensicherheit bis hin zu generellen Strategien für die digitale Wertschöpfung bei kleinen und mittelständischen Unternehmen und auch bei Großunternehmen.

Im Kontext dieser Richtlinie geht es bei Big Data um Technologien zur Datenanalyse. Entsprechende Algorithmen und Werkzeuge können Erkenntnisse über betriebliche Abläufe liefern und zu deren Optimierung beitragen. Hierzu bedarf es der Umsetzung dieser Methoden und Werkzeuge zur Verarbeitung, Analyse und Interpretation von umfangreichen und komplexen Daten in Big-Data-Anwendungen.

Die Richtlinienreihe unterstützt Erstellende und Nutzende bei der Vorbereitung, Entwicklung, Inbetriebnahme dieser Anwendungen und ihrem nachhaltigen Einsatz. Letztendlich sollen diese Big-Data-Anwendungen verlässlichere Entscheidungs-

a significant impact on the quality (or quantity) of the value creation of data projects. Since data-driven models are formulated with training data and checked with validation data, the quality of the model depends to a large extent on the quality of the training and validation data. In this context, the term “data quality” (see also Section 4) describes the characteristics of a dataset that make it suitable for fulfilling defined and presupposed requirements. Therefore, data quality can only be assessed in connection with data use and data generation, which are dealt with in detail in VDI/VDE 3714 Part 3.

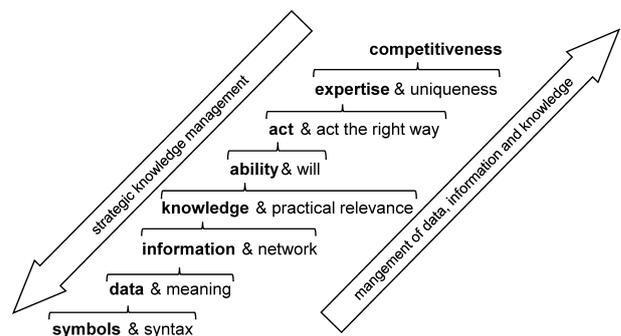


Figure 1. Knowledge staircase according to North

## 1 Scope

Although the term “big data” has been used for several years, it continues to be associated with very different topics and aspects and is accordingly discussed in a differentiated manner in the social debate. The ever-advancing digital communication, the broadband expansion that is currently being implemented, and the processing of data that is possible everywhere are fuelling this discussion both in the public and in the professional world. The topics range from data protection and data security to general strategies for digital value creation for small and medium-sized enterprises as well as for large companies.

In the context of this standard, big data is about data analysis technologies. Corresponding algorithms and tools can provide insights into operational processes and contribute to their optimization. This requires the implementation of these methods and tools for processing, analysing and interpreting extensive and complex data in big data applications.

The series of standards supports creators and users in the preparation, development, commissioning of these applications and their sustainable use. Ultimately, these big data applications should provide a more reliable basis for decision-making in order to

grundlagen schaffen, um Produkte und Produktionsprozesse ökonomisch, ökologisch und technisch zu verbessern.

Die Richtlinienreihe soll dazu beitragen, die Vielfalt der in den letzten Jahren durch Forschungs-, Entwicklungs- und Praxisarbeiten entstandenen Erkenntnisse aufzubereiten, die Entwicklung und den Einsatz von Big-Data-Anwendungen in produzierenden Industrien sowie deren Nutzung im regulären Betrieb zu unterstützen.

Zur Zielgruppe gehören alle Stakeholder, von den Praktikern/Praktikerinnen bis zu den Entscheidern/Entscheiderinnen, von der Fertigungs- bis zur Prozessindustrie. Die Richtlinienreihe wendet sich dabei an Nutzenden und Erstellenden von Big-Data-Anwendungen in der produzierenden Industrie, unabhängig und übergreifend für alle Führungs- und Fachaufgaben.

improve products and production processes economically, ecologically, and technically.

The series of standards is intended to help process the variety of findings that have emerged in recent years through research, development and practical work, and to support the development and use of big data applications in manufacturing industries as well as their use in regular operations.

The target audience includes all stakeholders, from practitioners to decision makers, from operations to process industries. In this context, the series of standards addresses users and creators of big data applications in the manufacturing industry, independently and across all management and technical tasks.