Towards integrated Data Analysis Quality: Criteria for the application of Industrial Data Science

Nikolai West
Technical University Dortmund
Institute of Production Systems
Dortmund, Germany
[0000-0002-3657-0211]

Jonas Gries
Technical University Kaiserslautern
Institute of Virtual Product Engineering
Kaiserslautern, Germany
[0000-0002-1504-1492]

Carina Brockmeier
Technical University Dortmund
Institute of Production Systems
Dortmund, Germany
[0000-0003-3495-4518]

Jens C. Göbel
Technical University
Kaiserslautern
Institute of Virtual Product
Engineering
Kaiserslautern, Germany

Jochen Deuse
Technical University Dortmund,
Institute of Production Systems
Dortmund, Germany;
Centre for Adv. Manufacturing,
University of Technology Sydney
Ultimo NSW, Australia
[0000-0003-4066-4357]

2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI) | 978-1-6654-3875-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/IRI51335.2021.00024

Abstract—**The application of Industrial Data Science in context of connected Smart Products requires modeling and structuring data for its design, development and use. Especially for Smart Products, a comprehensive handling of data quality is mandatory, because of their interdisciplinary character and broad range of heterogeneous stakeholders covering the entire product lifecycle. The overall goal of data preparation is to provide high-quality data for application and evaluation by users. Established process models for industrial data analysis often treat the specification and assurance of data quality as a single-point activity with a defined conclusion. Providing end-to-end data quality has received little attention in the field of industrial data analytics. In this paper, we will (1) structure four distinct phases for ensuring end-to-end data quality along data analytics activities, (2) define a set of criteria and measures for meeting and quantifying data quality requirements based on established criteria, and (3) provide a step-by-step model for establishing and maintaining high Data Quality for Industrial Data Science applications. The quality criteria aim to identify pointwise and continuous actions during the data analysis process. Such criteria target a shared responsibility for maintaining data quality during analyses between analyst and user. The developed model provides an actionable approach for assessing and ensuring the requirements of Data Analysis Quality.**

Keywords—Data quality, data analysis quality, industrial data science, quality management, quality assurance, data quality criteria

I. Introduction

Ever since the *First Industrial Revolution*, the manufacturing industry relies on quantitative and fact-based assessments for operational decisions and optimization measures. Thus, data has become an economic commodity and it takes an integral part of organizations. In a popular science article in The Economist, data is even considered the most valuable resource next to oil [1]. Low-cost hardware and open source platforms enable the collection and processing of the increasing volumes of data [2]. Products themselves become increasingly smart and create more data. Smart products are highly interdisciplinary with embedded intelligence, connectivity and offer reconfiguration during the product usage phase. A product digital twin based on development models enriched with data from the production and usage phase enables disruptively new data-driven-business models and opportunities [3]. Based on this data, products and related processes can be optimized to fit customer needs. Naturally, companies strive to capture and preserve data as economic assets. As volumes increase, it is essential to address issues of quality to derive meaningful insights from data.

Assessing the quality of data requires a context, as it can only be evaluated based on purpose and usage. Such context is often referred to as the *Fitness for Use*. Joseph M. Juran, whom many consider a pioneer of Quality Management, coined the term to define all aspects related to quality [4]. Fitness for Use encompasses the innumerable factors that define quality and has gained acceptance in both academic and industrial settings [5]. Although the concept of quality is still a matter of debate, the Fitness for Use has been the de facto definition of quality for over three decades [6]. As such, the concept is at the heart of the definition of *Data Quality* (DQ). DQ describes 'the degree to which data is fit for use by data consumers' [7]. To specify the diverse requirements, many researchers provided detailed sets of criteria to describe and measure individual aspects of DQ. This article reviews several of such criteria sets in detail below.

To apply *Industrial Data Science* (IDS), DQ and a consumer fitness for use is necessary. We refer to IDS as the use of Data Analytics in industrial settings and it serves as a tool for fact-based decision-making in value creation networks [8]. As a branch of Data Science, IDS addresses domain-specific analysis scenarios in particular [9]. Established process models for data analysis consider the assurance of DQ to be a necessary step during data preprocessing [10]. Similar to a *Quality Gate* in manufacturing, most models treat DQ as a single-point activity with a predetermined conclusion. After fulfilling the predefined DQ criteria, i.e. the passing of said Quality Gate, this approach assumes a lasting standard of DQ. However, it is commonly understood that, during the analysis, the object of consideration changes: Through industrial Data Analysis, *data* is linked to a

978-1-6654-3875-9/21/$31.00 ©2021 IEEE
DOI 10.1109/IRI51335.2021.00024

meaning, which is thereby becoming *information*. Linking the obtained information to a domain-specific context provides novel and potentially useful *knowledge* [11]. If the object of consideration changes, so must the Fitness for Use criteria. While in IDS the data is considered first, using it as information requires domain expertise and the subsequent use of the knowledge is inextricably bound to a use case. Most DQ criteria consider the suitability of data, while some specifically address *Information Quality*, but none account for the changing nature during on-going analysis. To comply with the Fitness for Use principle, we propose the concept of Data Analysis criteria as an integrated approach to ensure consistent DQ for IDS projects.

II. Fundamentals

A. Traditional Dimensions of Data Quality

Twenty-five years ago, Richard Wang and Diane Strong identified 15 dimensions for classifying and measuring data quality through two quantitative surveys and published them in the *Journal of Management Information Systems* [7]. Said publication takes into account the aforementioned Fitness for Use concept [4] and provides one of the earliest sets of data quality criteria. Consideration of the Fitness for Use led to a high emphasis on the consumer perspective for the quality criteria. Only the users of data can judge the suitability of the data quality for practical application purposes, hence their perspective is the main focus of consideration [7]. Since it is only during the operational usage of data that an economic benefit arises, this perspective may have contributed significantly to the subsequent success of the publication. Initially, Wang and Strong based their criteria for data quality characteristics on the mentioned surveys, which comprised a set of 179 individual features. By evaluating the relevance of this wide range of characteristics, the authors were able to narrow their set down to 20 data quality criteria. Since 20 dimensions were too many for practical evaluation purposes, they first had to divide the criteria into target categories. This allows adjustments regarding the necessity of individual criteria in their respective category. As a result, five criteria were eliminated as well, leaving the final set of 15 unique criteria for data quality, listed in TABLE I.

TABLE I. Wang and Strong's Framework of Data Quality [7]

Category	Dimension	
Intrinsic Data Quality	Believability	Objectivity
	Accuracy	Reputation
Contextual Data Quality	Value-added	Timeliness
	Relevancy	Completeness
	Appropriate amount of data	
Representational Data Quality	Interpretability	Representational consistency
	Ease of understanding	Concise representation
Accessibility Data Quality	Accessibility	Access security

This conceptual framework addresses the multi-dimensional nature of DQ and provides a division of responsibilities through the categorization of the criteria. *Intrinsic Data Quality* implies that data has a quality in itself. *Contextual Data Quality* emphasizes the requirement that DQ requires consideration in the context of a particular task. *Representational Data Quality* and *Accessibility Data Quality* emphasize the importance and influence of the systems under consideration. In a follow-up publication, the authors demonstrated the suitability of the dimensions as well as the categories using three case studies [12]. We consider a selection of approaches in the next section.

B. Approaches to categorize Data Quality

Although the dimensions of Wang and Strong are still widely used today, multiple other approaches provide criteria-based descriptions for data quality. One such set of DQ criteria stems from the *Data Administration Management Association* (DAMA), a non-profit and vendor-independent association dedicated to the advancement of data and information resource management [13]. In 2013, a working group of DAMA published a concept with condensed criteria for DQ [14]. Aiming to define best practice definitions of generic data quality dimensions, they postulate the following six criteria:

- Accuracy
- Completeness
- Consistency
- Timeliness
- Uniqueness
- Validity

Focusing on these six core dimensions aims to enable crucial understanding and management of data. Using this simplified approach, organizations select data quality dimensions and associated dimension thresholds based on individual demands, such as business context, technical requirements or a risk level. In the process, each dimension gets an individual weighting. To obtain an accurate measure of data quality, the working group suggests that an organization must determine how much each dimension contributes to data quality as a whole. Additionally, to ensure an effective use of data, they recommend that other factors for consideration along the six dimensions, like the *usability* of the data, *timing issues* with the data, *flexibility* of the data, *confidence* in the data, and *value* of the data will be reviewed [14]. Although these six core dimensions do not have the scope of Wang and Strong's framework, they allow for a summarized consideration of important aspects of quality.

Another set of criteria resulted from a similar context: A working group of the *German Association for Information and Data Quality* (Deutsche Gesellschaft für Informationsqualität, DGIQ) derived 15 dimensions of DQ, which are divided into four categories [15]. The criteria provided there have a direct relation to the original dimensions of Wang and Strong, but differ in two aspects. Firstly, the DGIQ does not consider *access security* to be an independent dimension, but rather a limitation, that must be defined together with other dimensions. Secondly, the DGQI includes the *ease of manipulation* as an additional DQ dimension. The scientific origin of this criterion also traces back to Wang, who has included it among the DQ dimensions in a subsequent publication [16]. Besides these minor adjustments, the collection of criteria remains largely unchanged twenty years after the original publication *(see Fig. 1)*. DGIQ also retains the original categorical division of the criteria, but renames them using the following terms for ordering: *system-supported*, *purpose-dependent*, *data-inherent* and *representation-related*. A novelty of the DGIQ criteria is the shift in their field of focus. While WANG and STRONG refer primarily to DQ, the DGIQ designates the criteria as dimensions for *Information Quality* (IQ). By defining these dimensions, DGIQ has succeeded in simplifying and improving communication on DQ and IQ management topics by using clear and uniform terminology in German-speaking countries. In addition, the DGIQ's approaches to quantify the effects of quality issues made an important contribution to the assessment and measurement of DQ or IQ.

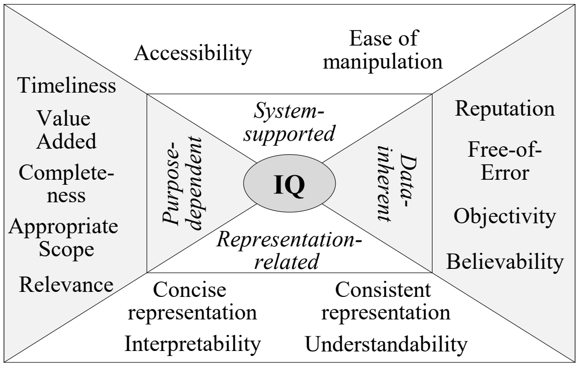

Fig. 1. Translated DGIQ criteria for Information Quality [15]

Categorization of data quality requirements is a popular subject of research. The concepts of DAMA and DGIQ are merely a selection of available approaches. For a comprehensive overview we refer to designated research literature [6, 17–19].

C. Demand for Data Quality in Industrial Data Science

Due to increasing scale and complexity, manual methods for data processing are no longer economical. Thus, manufacturing companies seek the use of IDS for the efficient evaluation and utilization of implicitly available knowledge [20]. As for *Knowledge Discovery in Databases*, IDS includes all non-trivial measures to identify valid, novel, potentially useful, and ultimately understandable patterns in industrial data sets [21]. Typically, process models are used to implement IDS projects, with the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) being one of the most common approaches [10]. *Fig. 2* shows the six iterative process steps of the CRISP-DM: *Business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, and *deployment*. In accordance with the iterative process character, process steps are typically run through one after the other, but the model allows for individual backward steps or renewed iterations. Due to its successful dissemination, the CRISP-DM is considered one of the most common process models in IDS and it serves as the organizational basis for the considerations made in Chap. III.

When applying quality criteria to IDS projects, the multi-dimensional criteria sets must consider the lifecycle-dependent nature of their observation target. We mentioned this change of the object of consideration in the introduction. In the Data Analytics context, this refers to the shift from *data* over *information* to *knowledge*. According to the so-called *Staircase of Knowledge,* data inevitably passes through a number of stages during commercial use [11]. Characters with a syntax represent raw data. Information arises through the linkage with a real-world meaning and knowledge is then obtained by relating said information to a domain-specific context. In addition, the Staircase of Knowledge presents the subsequent formation of skills, actions, competencies and ultimately competitiveness.

For applied DQ criteria, we limit our considerations to data, information and knowledge, which we integrated into the scope of the CRISP-DM in Fig. 2. The second step of the CRISP-DM, *Data Understanding*, involves the verification of DQ. For this purpose, the CRISP-DM offers exemplary guiding questions and refers to the requirements regarding the necessary quality of data and results that were prior assessed in the phase *Business Understanding*. This involves creating a data quality report that becomes the focus for data cleaning in the next phase, *Data Preparation*. After this point, the CRISP-DM considers DQ to be guaranteed in general and is only briefly revisited during *Evaluation* of the process results [10]. For use in a higher-level process model, we consider this approach appropriate. However, considering the changing nature of data during IDS projects, we regard it as necessary to realign the dimensions or criteria of DQ.

A similar concept has already been followed with the *Total Data Quality Management* (TDQM) [16]. The TDQM aims to extend traditional *Total Quality Management* that demands for a consideration of product quality over the entire life cycle. We refer to a process-oriented management of all aspects related to the quality of data, information and knowledge during applied IDS as *Data Analysis Quality* (DAQ) and define it as follows:

Data Analysis Quality describes the extent to which data products are fit for use by Industrial Data Science consumers.

In the following chapter, we will introduce a set of criteria for DAQ and structure it in a process chain for applying IDS.

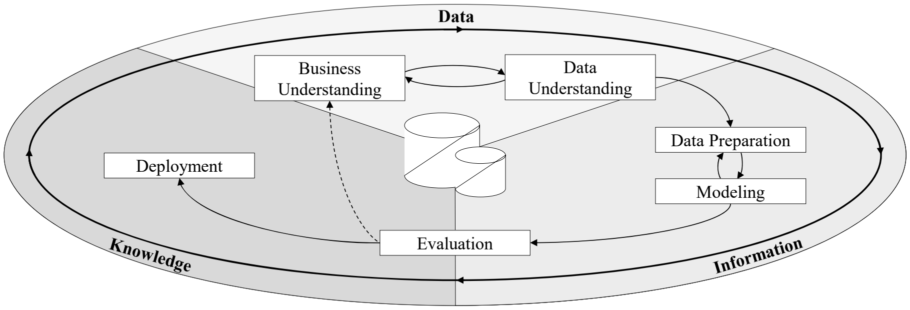

Fig. 2. Integrated visualization of the *Cross Industry Standard Process for Data Mining* [10] and a selection of steps from the *Staircase of Knowledge* [11]

III. Integrated Data Analysis Quality

A. Phases to ensure end-to-end Data Analysis Quality

Typically, collections of DQ criteria include a grouping of the requirements to facilitate their utilization. We similarly classify the proposed DAQ criteria to account for the changed perspective over the course of data usage. Since the DAQ aims to provide an industrial perspective on DQ, we build the categories on an observation made in *Industrial Engineering* (IE). In most manufacturing companies, IE provides system-, method- and problem-solving competences along the evolving requirements of the continuous improvement process [22]. As such, IE and IDS are closely related, since both involve similar steps for fact-based decision-making processes *(see Fig. 3)*.

1. Since fact-based decisions require a quantifiable basis, the first step is to ***access*** all necessary data sources. This also includes identifying relevant sources for a given analysis, recoding missing data with suitable collection methods, and providing the data to an analyzing system as described in [23]. Thus, the step relates to the initial phases in the CRISP-DM, *Business Understanding* and *Data Understanding*. It deals primarily with the access and provision of *data* in an unprocessed form for all subsequent analysis steps.

2. After ensuring end-to-end access to the relevant data, the second step is to ***analyze*** the data to obtain *information*. Just like IE, IDS can draw on a broad selection of potential tools and methods, such as the myriad of open source packages and data science algorithms. Additional steps for information processing and transformation for subsequent usage may accompany the analysis step. In the CRISP-DM, it affects the enclosed steps *Data Understanding* and *Evaluation*, but forges *Data Preparation* and *Modeling*.

3. Economic benefit only arises through the operational use of the information generated during the analysis. In the third step, it is therefore necessary to ***apply*** the information in an industrial use case. This includes both the realization of selective analyses for targeted questions as well as the implementation of continuous evaluations of long-term observations. In line with CRISP-DM, this is *Deployment*, in which problem- and domain-specific knowledge helps to gain a monetary value from the overall IDS project.

4. Although these steps generally conclude an analysis, the industrial reality indicates the need for a fourth step to ***administrate*** the wealth of peripheral processes. The step may not have a direct pendant in the CRISP-DM, but it is eligible as collective pool for indirect operational and organizational tasks that relate to the analysis process. This may include tasks such as assigning a long-term data stewardship, allocating a data governance, ensuring an end-to-end data security and securing an ethical data usage.

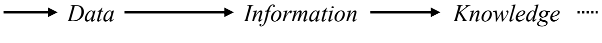

Fig. 3. Four steps for categorizing the criteria of Data Analysis Quality

Using these four steps, we will structure the proposed DAQ criteria. The criteria are assigned to the earliest applicable step and therefore be valid in the ongoing process. In case of later revealed quality problems, iterations have to be possible. Ensuring an integrated DAQ then adheres to this process sequence. Throughout this paper, we refer to the steps as *layers*. For ease of reading, we synonymously use the term data for information and knowledge during criteria description.

B. Criteria catalog for Data Analysis Quality

The focus of consideration in the following section is on the individual criteria used to evaluate and measure DAQ. Since linguistic ambiguities in criteria terminology can contribute to misinterpretations in their use, we put the primary focus on an explanation of the objective of each criterion. All but two of the DAQ criteria rely, directly or indirectly, on prior scientific work by other authors. We provide references to the sources during the criteria descriptions. The secondary focus of our elaborations lies on a quantification of the DAQ criteria, expressed by a measurable degree. We provide some considerations for the evaluation of quality, but leave the use-case-based specification to future practitioners, due to the extensive context dependency. The following criteria provide a broad overview of the different perspectives on DQ and represent a comprehensive framework.

1) Access layer

The first layer covers the gathering of needed data according to the defined analysis goals of the *Business Understanding* phase. Following specifications of *Data Understanding*, the layer treats further aspects regarding the quality of raw data and the corresponding business processes *(see Fig. 2)*. The criteria shown in TABLE II. support the tasks in the following layers.

TABLE II. Criteria to access data

Criteria	Definition	Sources
Accessibility	The extent to which data is readily available or quickly retrievable	[7, 15]
Relevancy	The extent to which data is needed and applicable for the given task	[7, 15]
Timeliness	The extent to which data is accessible in a required time	[24, 25]
Uniqueness	The extent to which data measures events or objects no more than once	[14, 26]
Validity	The extent to which data conforms to a predefined set of rules	[14, 25]

Accessibility. The criterion *'accessibility'* demands that the data of interest can be repeatedly obtained from a data source through defined interfaces [15]. Accessibility covers providing data for the downstream processes and therefore facilitates the modifiability of data. In the following step of *Data Preparation*, the data is accessed and local copies of the data being edited and transformed to fit the requirements of the analysis. In a DAQ context, it makes Wang and Strong's criterion *'ease of manipulation'* redundant [7, 16]. Problems may result from missing rights of data ownership or inadequate process-related data acquisition. We consider it a fundamental prerequisite to perform any kind of data analysis. For measurement, a binary value, accessible or non-accessible, is sufficient.

RELEVANCY. The criterion *'relevancy'* assesses the usefulness of the data for the planned use case [7, 15]. The scope and amount of data, which will be provided to the phase of *Data Preparation*, has to fit according to the predefined analysis target. Therefore, the extent of data has to cover all required data but not include unnecessary data. This would enlarge runtimes and cause extra work in the upcoming phases. The criterion is linked to *completeness*, whereas here the focus is on the extent of the data set, while it later focuses the attention on a contextual completeness. To measure the criterion, we propose three ordinal characteristics: insufficient, ideal or excessive.

TIMELINESS. The criterion *'timeliness'*, which may also be called latency, focusses on the duration necessary to access data [24, 25]. The access duration is dependent on the technology used for storing and retrieving the industrial data, covering both software and hardware. There may also be a dependency on processual aspects, such as the capability to collect the required process data in time. Similar to the *accessibility*, there is just a necessity for a binary metric: in time or not in time.

UNIQUENESS. The *'uniqueness'* of data originally describes an absence of redundancy [26]. Accordingly, every set of data describing a single event or object owns a unique key [26]. However, since IDS projects often require data from more than one data source, a comprehensive uniqueness is often hard to achieve. Varying data sources may contain different types of data relating to the same event or object. To create a coherent view of the data, a systemic approach to data integration is necessary. The extent of integration covers a broad variety differing in complexity and use, whereby the criterion of *uniqueness* in the context of DAQ requires a basic form of integration, targeting a freeness of technical duplicates. A more powerful, semantic integration resolving ambiguities is more appropriately situated in the following step of *Data Preparation*.

VALIDITY. The criterion *'validity'* rates the conformity of data to given rules [25]. Such rules may be of normative nature, like a file format specification, or of non-normative nature, like physical boundary conditions for value intervals [25]. Different to *free-of-error*, these rules are always applicable for the defined type of data and thus not considered use case specific. The criterion thereby covers a technical interpretability of the data, related to the traditional understanding of *interpretability* [7]. Please note the different interpretation in the application layer. The validity rules can be checked with the help of automated routines , due to their general applicability. Like the previous criteria, only a binary decision for every defined rule is required.

2) Analysis layer

The second layer relates to the quality of the data analysis and primarily deals with *information*. It corresponds to the second and third step of the CRISP-DM, *Preparation* and *Modeling*. During this phase, the accessed data needs the context of a dedicated use case or a problem statement. The context guides the processing steps using at least one dedicated analysis method. TABLE III. summarizes the criteria for the analysis.

ACCURACY. The criterion *'accuracy'* defines the degree to which data correctly describes a real-world object or event [14, 24]. As such, we determine *accuracy* by the extent to which values match a defined reference source of correct observations. It requires the comparison of values from a given source of truth against the dynamically calculated values in the manufacturing system. Alternatively, if no such source is available, manual verification is a valid option to confirm the data accuracy [26]. Accuracy is the quotient of the number of correct values in a source and the total number of values in the source [27]. In the use of data for analysis, there are task-dependent requirements for the acceptable degree of deviation. These deviations require consideration of business expectations and problem definition. For IDS, it is crucial to reach the needed degree of accuracy.

TABLE III. CRITERIA IN THE ANALYSIS LAYER

Criteria	Definition	Sources
Accuracy	The extent to which data correctly describes an object or event at hand	[7, 14, 27]
Completeness	The extent to which data is not missing and of sufficient scope	[14, 26, 28]
Free-of-error	The extent to which the data provided is correct and reliable	[15, 28]
Value-added	The extent to which data allows benefits from its analytics use	[7, 12, 15, 28]

COMPLETENESS. According to the criterion *'completeness'*, data is complete if they are not missing and are available at the specified times in the respective process steps [26]. The criterion relates closely to *relevancy* and *timeliness*. The extent to which the data has sufficient breadth, depth, and scope is explicitly dependent on the task at hand [7]. The concept of completeness implies the existence of non-zero values associated with specific data elements [24]. Thus, completeness can be determined as a quotient composed of the number of non-zero values in a source and total size of the data source [27]. While non-zero values may be overcome using appropriate interpolation approaches, they might also represent termination conditions in other cases. Thus, the degree of completeness relates closely to the analysis task.

FREE-OF-ERROR. The criterion *'free-of-error'* relates closely to *accuracy*. While accuracy was originally included in WANG and STRONG's dimensions [7], later considerations substituted it for the criterion free-of-error [15, 28]. For DAQ, *accuracy* refers to permissible deviation tolerances. *Free-of-errors* instead intends to ensure the absence of any logical inconsistencies. Typically, a use case determines the plausibility of values. This may include restrictions of a value range or logical conditions. By guaranteeing that data is free of error, DAQ ensures that the data analysis uses and passes only plausible information.

VALUE-ADDED. According to the *'value-added'* criterion, DAQ considers value creating if its usage can lead to a quantifiable increase in novel information. While data may be processed in numerous ways, only a value-adding use advances the data analytics objectives. This describes the extent to which data has utility and allows advantages from its use [7]. This criterion is typically required in decision-support information systems for which cost-benefit calculations are performed. Once again, the metric results in a quotient derived from the monetary outlay for the solution as well as the added values achieved [27].

3) Application layer

The third layer addresses the application of data in an industrial setting. As such, this phase also has a reference in the CRISP-DM with the *Evaluation*, as well as a direct counterpart with the *Deployment*. The criteria in this layer aim to establish a high quality for the results of data analyses. They specifically target the inclusion of personnel not previously involved in the analysis. As future users of the deployment solutions, the level of fit represents the overarching goal at this stage. TABLE IV. outlines the criteria for the application layer.

TABLE IV. CRITERIA IN THE APPLICATION LAYER

Criteria	Definition	Sources
Cost-effectiveness	The extent to which the analysis' benefits cover implementation costs	[7]
Concise representation	The extent to which data products are presented in a compact format	[7, 28]
Consistent representation	The extent to which data products are presented in a unified format	[14, 28]
Inter-pretability	The extent to which data products address the context of consumers	[7, 28]
Under-standability	The extent to which data products allow for actionably statements	[7, 28]

COST-EFFECTIVENESS. While the *'cost-effectiveness'* was eliminated in the publication by WANG and STRONG due to lack of relevance [7], we propagate a use of the criterion for DAQ criteria. Unlike the traditional understanding, we refer not only to the costs of data collection and storage, but also to the efforts to utilize the data analytic solution in the enterprise. As a criterion, the *cost-effectiveness* relates closely to the criterion *value-added* from the previous layer. While *value-added* seeks to assess the potential value of analysis results, *cost-effectiveness* includes both the achieved value and the costs incurred through implementation. The determination can be made either after the introduction of the data-analytical solution in everyday operations, or in advance with the help of suitable estimate calculations. From an economic point of view, we consider the criterion of outstanding importance in IDS projects.

CONCISE REPRESENTATION. The criterion *'concise representation'* indicates the extent to which analysis results are presented in a compact form without being overpowered by excessive complexity [7]. It targets the usage of deployment solutions that are brief in presentation but convey their core message completely and to the point. In particular, as models become more complicated and require a representation as black box models, the resulting representations must be unequivocal and unambiguous to enable the ability to act accordingly.

CONSISTENT REPRESENTATION. The intent behind the *'consistent representation'* criterion is similar to the previous criterion *concise representation*. The use of knowledge as a prepared deployment solution intends to support the best possible use in operation. The solutions must therefore be as uniform as possible in order to maintain a consistent understanding of the methods in changing use cases. Therefore, this criterion no longer just describes the extent to which data is presented in the same format and compatible with previous data [7]. For application as a criterion for the quality of data analysis, we extend the understanding of such uniformity to refer explicitly to the presentation manner of the developed analysis solutions. The consistency of representation is measurable as the degree to which the structure of solutions conforms to the common standard of all representations.

INTERPRETABILITY. The *'interpretability'* of the developed solutions is crucial for the exploitation of obtained information. The criterion measures the extent to which the products of the IDS project are in appropriate languages, symbols and units with defined definitions [28]. This measurement refers primarily to the aspects of the chosen presentation method, which deal with the utilization of analysis results by data consumers. Instead of a quantitative assessment of interpretability, we propose a qualitative assessment of IDS products using employee surveys.

UNDERSTANDABILITY. Despite the linguistic similarity to the previous criteria, the *'understandability'* offers a novel perspective for DAQ by providing implied immediate feedback from knowledge consumers. WANG and STRONG defined the *ease of understanding* as the extent to which data are clear without ambiguity and easily comprehended [7]. In the context of DAQ, we consider *understandability* as a measure of the ability to make operational decisions from processed information. It denotes the amount of knowledge acquired from an IDS product and is measurable using qualitative interviews.

4) Administration layer

The final layer addresses issues of data administration. TABLE V. shows the proposed criteria that support the previous layers and are relevant for all phases of the CRISP-DM.

TABLE V. CRITERIA TO ADMINISTRATE DATA

Criteria	Definition	Sources
Security	The extent to which access to data is restricted appropriately to maintain its security	[7, 27, 29]
Verifiability	The extent to which the data correctness and trustworthiness can be determined by defined activities	[27, 30]
Confidenti-ality	The extent to which confidential data is sufficiently protected	[29]

SECURITY. The criterion *'security'* demands the use of technically secured data to prevent misuse or manipulation [7]. Therefore, the data source should offer measures like an authentication mechanism and the data transmission should use suitable encryption, especially if the transmission is using public telecommunication networks. The effort spend in security measures depends on the value that the data is carrying and is thus highly varying by company [27].

Verifiability. The criterion 'verifiability' addresses the trustworthiness and correctness of data [27]. Data for IDS often originates from various sources, e.g. automatically collected data, human generated data or data delivered by suppliers or partners. To ensure the correctness and trustworthiness, such data requires verification. When planning verification activities, the former reliability of the data source can be an indicator for the needed expenditure. In conjunction with *validity*, this criterion ensures the *believability*, as advocated by Wang and Strong in their original dimensions for DQ [7].

Confidentiality. The criterion *'confidentiality'* assesses the protection requirements of sensitive data [29]. The sensitivity of data can be determined either by their potential business value or by statutory and social restriction. Examples for the business value are product or process specific data describing unique features or technologies used. The handling of personal data for instance, is restricted statutory regulations and, especially in Germany, demands special protection.

C. Process model for integrated Data Analysis Quality

As elaborated in the introduction, we consider the assurance of integrated quality during the steps of industrial data analyses to be an end-to-end task. Instead of single-point activities, DAQ requires a continuous process that covers the lifecycle of data, information and knowledge. With the division into four layers, which are oriented according to the characteristics of the object under consideration, the DAQ criteria enable a stepwise and task-oriented approach. A synchronized application of an IDS project with the steps of CRISP-DM requires a process model that allows iteration and back stepping. Such a model must define the sequence of the four layers and serves as the basis for a distribution of operational responsibility. Fig. 4 shows the proposed model that incorporates the four DAQ layers *(see Fig. 3)* and includes the criteria from TABLE II. to TABLE V. The model is not a replacement for the CRISP-DM, but a supplement for ensuring integrated DAQ. While the model suggests to successively progress through the criteria in the *Access Layer*, *Analysis Layer*, and *Application Layer*, the criteria in the *Administration Layer* must be ensured in an ongoing effort. To reach the second and third layer, a use case must fulfill the criteria from previous layers. In case of project adjustments or unexpected disruptions during the IDS project, backward steps are possible at any time. By adhering to all criteria in the process model, integrated data quality is achieved in the IDS project. We call the collection of criteria generally applicable, but make no claim to universal exhaustiveness. Due to the highly individual requirements of IDS projects, an adaptation or extension of the criteria selection may be necessary. The respective degree of fulfillment of the criteria is in any case application-dependent.

IV. Conclusion

This paper contributes to the development of integrated *Data Analysis Quality* for the application of *Industrial Data Science*. Based on established approaches to categorize DQ, we presented an organizational approach using four layers in the IDS process. Our considerations build on the traditional approaches to categorize DQ, such as the popular dimensions of Wang and Strong [7] as well as the approaches of DAMA [14] and DGIQ [15]. Each of the proposed layer contains a subset of criteria to ensure and evaluating the state of DAQ. They explicitly align with the current state of the analysis object during each layer. This allows an improved consideration of the fitness of data, information, and knowledge during the process of the industrial data analysis. As noted in the proposed definition, DAQ thus benefits all consumers during the IDS process. This includes potential consumer groups from different layers, such as data backend engineers, data analysts as well as domain users. Given the operational necessity, we have extended the notion of DQ by defining DAQ for IDS projects. Due to a methodological resemblance to the widely used CRISP-DM, we ensure a high practicability of the developed DAQ criteria. The process model serves as a complement to CRISP-DM and improves the handling of DQ, which many still treat as a single-point activity.

Overall, the framework, in conjunction with the criteria, enables the realization of a holistic DQ strategy such as TDQM. However, it is important to initiate suitable measures to ensure a sustainable DAQ over the entire product or process lifecycle and if quality aspects are monitored on the long term. Creating an appropriate mindset and awareness for DQ, similar to the field of production, which is highly influenced by the ISO 9001, is essential to sustainably ensuring a proper DQ. The theoretical description of the criteria laid the foundation for the future use of DAQ. As a following step, we plan to validate the model within different use cases from the research project AKKORD *(see Acknowledgement)*. The methodology will be a part of the reference building block for industrial data analysis, which will enable easier application of data analysis even for small and medium-sized enterprises. A deliberate handling of data quality from start to end will ensure more efficient and successful analysis projects. In summary, a networked and integrated application of industrial data analytics for value-creating, competency-based collaboration in dynamic value networks requires a holistic approach [8]. Regardless of the criteria or dimensions ultimately chosen, it is essential to address the changing nature of the object of consideration, from available *data* to novel *information* to value-added *knowledge*.

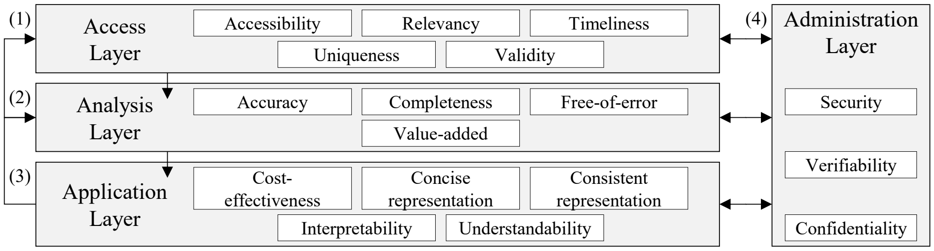

Fig. 4. Process model for integrated Data Analytics Quality over the course of the four layers Access, Analysis, Application and Administration

ACKNOWLEDGMENT

The work on this paper has been supported by the German Federal Ministry of Education and Research (BMBF) as part of the funding program 'Industry 4.0 - Collaborations in Dynamic Value Networks (InKoWe)' in the research project AKKORD (02P17D210); *www.akkord-projekt.de*.

V. REFERENCES

[1] The Economist, *The world's most valuable resource is no longer oil, but data.* [Online]. Available: economist.com/ leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (accessed: Jul. 1 2021).

[2] R. Wöstmann, A. Barthelmey, N. West, and J. Deuse, "A Retrofit Approach for Predictive Maintenance," *Tagungsband des Kongresses Montage Handhabung Industrieroboter*, vol. 4, no. 1, pp. 94–106, 2019.

[3] J. C. Göbel and T. Eickhoff, "Conception of digital twins for smart products (in german)," *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, vol. 115, s1, pp. 74–77, 2020, doi: 10.3139/104.112301.

[4] J. M. Juran and F. M. Gryna, *Quality Planning and Analysis: From Product Development through Use,* 3rd ed. New York: McGraw-Hill, 1993.

[5] M. J. Eppler, *Managing Information Quality*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[6] M. P. Neely, "The Product Approach to Data Quality and Fitness for Use: A Framework for Analysis," *Proceedings of the International Conference on Information Quality*, vol. 10, no. 1, pp. 52–66, 2005.

[7] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–34, 1996.

[8] J. Mazarov, P. Wolf, J. Schallow, J. Deuse, and R. Richter, "Industrial Data Science in Value Creation Networks (in German)," *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, vol. 114, no. 12, pp. 874–877, 2019.

[9] N. Bauer *et al.,* "Industrial Data Science: Developing a Qualification Concept for Machine Learning in Industrial Production," *Archives of Data Science, Series A*, vol. 5, no. 1, pp. 1–14, 2018.

[10] P. Chapman *et al., CRISP-DM 1.0: Step-by-step Data Mining Guide*: CRISP-DM Consortium, 1999.

[11] K. North, *Knowledge-based Management: Creating Value through Knowledge (in German),* 5th ed. Wiesbaden: Springer Gabler, 2011.

[12] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data Quality in Context," *Communications of the ACM*, vol. 40, no. 5, pp. 103–110, 1997.

[13] D. Henderson and S. Earley, *DAMA-DMBOK: Data Management Body of Knowledge,* 2nd ed. Basking Ridge, NJ: DAMA, Technics Publications, 2017.

[14] N. Askham *et al., The Six Primary Dimensions for Data Quality Assessment: Defining Data Quality Dimensions*. Bristol: Data Management Association (UK), 2013.

[15] J. P. Rohweder, G. Kasten, D. Malzahn, A. Piro, and J. Schmid, "Information Quality: Definitions, dimensions and terms (in German)," in *Daten- und Informationsqualität*, K. Hildebrand, M. Gebauer, H. Hinrichs, and M. Mielke, Eds., Wiesbaden: Springer Vieweg, 2008, pp. 25–45.

[16] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, no. 2, pp. 58–65, 1998.

[17] A. Ramasamy and S. Chowdhury, "Big Data Quality Dimensions: A Systemic Literature Review," *Journal of Information Systems and Technology Management*, vol. 17, no. 1, 2020.

[18] F. Sidi, P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data Quality: A Survey of Data Quality Dimensions," *International Conference on Information Retrieval and Knowledge Management*, no. 1, pp. 300–304, 2012.

[19] M. Mirzaie, B. Behkamal, and S. Paydar, "Big Data Quality: A systematic literature review and future research directions," *arXiv: Databases*, pp. 1–12, 2019.

[20] M. Eickelmann, M. Wiegand, B. Konrad, and J. Deuse, "The Importance of Data Mining in the Context of Industry 4.0 (in German)," *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, vol. 110, no. 11, pp. 738–743, 2015.

[21] U. Fayyad, G. Piatetski-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, vol. 2, no. 1, pp. 82–88, 1996.

[22] R. Richter and J. Deuse, "Industrial Engineering in Modern Production (in German)," *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, vol. 207, no. 1, pp. 6–13, 2011.

[23] A. Eiden, J. Gries, T. Eickhoff, and J. C. Göbel, "Requirements for a data backend system to support industrial data analysis applications in digital engineering processes of dynamic value networks (in german)," *DFX-Symposium*, vol. 31, no. 1, pp. 81–90, 2020.

[24] D. Loshin, *Master Data Management,* 2nd ed. Amsterdam: Morgan Kaufmann, 2008.

[25] L. Sebastian-Coleman, *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Burlington: Elsevier Science, 2012.

[26] M. Mosley, M. Brackett, and S. Earley, *The DAMA guide to the data management body of knowledge*. Bradley Beach, NJ: Technics Publications LLC, 2010.

[27] F. Naumann, *Quality-Driven Query Answering for Integrated Information Systems*. Berlin, Heidelberg: Springer-Verlag, 2002.

[28] L. L. Pipino, Lee, Yang, W., and Wang, Richard, Y., "Data Quality Assessment," *Communications of the Association for Computing Machinery*, vol. 45, no. 4, pp. 211–218, 2002.

[29] D. E. Denning, *Cryptography and data security*. Reading, Mass.: Addison-Wesley, 1992.

[30] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith, "A Framework for Information Quality Assessment," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1720–1733, 2007.