D3.1 – Data Quality Guidelines

WP3 – BUILD: Manufacturing Data Quality

i4Q has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958205.

Document Information

GRANT AGREEMENT NUMBER	958205	ACRONYM	i4Q	
FULL TITLE	Industrial Data Services for Quality Control in Smart Manufacturing			
START DATE	01-01-2021	DURATION	36 months	
PROJECT URL	https://www.i4q-project.eu/			
DELIVERABLE	D3.1 – Data Quality Guidelines			
WORK PACKAGE	WP3 – BUILD: Manufacturing Data Quality			
DATE OF DELIVERY	CONTRACTUAL	June 2022	ACTUAL	June 2022
NATURE	Report	DISSEMINATION LEVEL	Public	
LEAD BENEFICIARY	BIBA			
RESPONSIBLE AUTHOR	Stefan Wellsandt (BIBA)			
CONTRIBUTIONS FROM	UNI, ITI			
TARGET AUDIENCE	1) i4Q Project partners; 2) industrial community; 3) other H2020 funded projects; 4) scientific community 5) manufacturing stakeholders with influence on quality management, like quality managers, researchers, and production managers.			
DELIVERABLE CONTEXT/ DEPENDENCIES	This document is the first version of the data quality guideline. Its relationship to other documents is as follows: • D3.9 Data Quality Guidelines v2 (M24): the revision of this deliverable with updates and extended examples • D3.2 QualiExplore for Data Quality Factor Knowledge: digital tool related to this guideline's suggested activities			
EXTERNAL ANNEXES/ SUPPORTING DOCUMENTS	None			
READING NOTES	None			
ABSTRACT	This deliverable contains a guideline to manage data quality in manufacturing. It establishes a conceptual basis by introducing several concepts, such as data and information, data life cycle, information needs, data and information quality, and production system levels. The guideline uses the Plan-Do-Study-Act (PDSA) cycle and focuses on the Plan and Do steps. It outlines an information flow analysis for producers to understand which data quality factors the organization must manage. Furthermore, it suggests three types of measures to manage data quality factors. Awareness measures aim to raise awareness of data quality issues and factors among employees.			

	They require the least effort but are also not very reliable unless strictly controlled. Programmatic measures are functions in software that force users into behavior that ensures high data quality. Examples are input validations and auto-complete. These measures are much more reliable but may be costly to implement. Organizational measures cover complex cases where other measures are not feasible. They focus on larger-scale organizational activities (e.g., work instructions, training, and new roles) to promote behavior that minimizes data quality issues.

Document History

VERSION	ISSUE DATE	STAGE	DESCRIPTION	CONTRIBUTOR
0.1	02-May-2022	ToC	Create ToC and Section Description	BIBA
0.2	31-May-2022	1st Draft	1st Draft available for review	BIBA
0.3	10-Jun-2022	Internal Review	Internal Review Process	IBM, ENG
0.4	20-Jun-2022	2nd Draft	Process comments from internal review	BIBA
1.0	30-Jun-2022	Final Draft	Final quality check and issue of final document	CERTH

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS/ACRONYMS

AI	Artificial intelligence
DQM	Data Quality Management
EDQ	Evolutional Data Quality
HCI	Human-computer interfaces
IM	Information Management
IQM	Information Quality Management
ISO	International Organization for Standardization
IT	Information Technologies
ML	Machine learning
PDSA	Plan-Do-Study-Act
WP	Work package

Executive summary

i4Q is, amongst others, devoted to providing methodologies, tools, and infrastructure to ensure high data quality in production. Meeting this objective will contribute to improved operational intelligence and data analysis results. Manufacturing data quality also ensures the needed accuracy and reliability of the data measured along the value chain. Data quality in manufacturing boosts (i) product quality in the supply chain; and (ii) process quality of the manufacturing companies. Data Quality in i4Q includes systematically identifying the factors that influence data quality in manufacturing by using data quality management and technologies supporting it.

This deliverable contains a **guideline to manage data quality in manufacturing**. It establishes a conceptual basis by introducing several concepts, such as data and information, data life cycle, information needs, data and information quality, and production system levels. The guideline uses the Plan-Do-Study-Act (PDSA) cycle and focuses on the Plan and Do steps. Section 3.1 outlines an information flow analysis for producers to understand which data quality factors the organization must manage. Section 3.2 suggests three types of measures to manage data quality factors. Awareness measures aim to raise awareness of data quality issues and factors among employees. They require the least effort but are also not very reliable unless strictly controlled. Programmatic measures are functions in software that force users into behavior that ensures high data quality. Examples are input validations and auto-complete. These measures are much more reliable but may be costly to implement. Organizational measures cover complex cases where other measures are not feasible. They focus on larger-scale organizational activities (e.g., work instructions, training, and new roles) to promote behavior that minimizes data quality issues.

The proposed activity framework in Section 3fulfills the first goal for task 3.1, i.e., providing key activities to manage data quality in manufacturing. This deliverable's second version will revise the activities based on insights from the use cases. The second goal concerns creating a methodological connection to other tasks (mainly in WP3). This document does not yet meet this goal, but the second version of this deliverable will contain an additional section to explain how other i4Q solutions and the activity framework align.

Document structure

Section 1: Contains a general description of the i4Q **Data Quality Guideline**, providing an overview and its goals. It is addressed to final users of this guideline.

Section 2: Contains the guidelines conceptual basis, relation to standards, and the overall suggested data quality management process. It is addressed to final users of this guideline.

Section 3: Details the i4Q **Data Quality Guideline**, explaining relevant activities. It is addressed to final users of this guideline.

Section 4: Provides the conclusions.

1. Introduction

1.1 Overview

i4Q is devoted to providing methodologies, tools, and infrastructure to ensure high data quality in production. Meeting this objective will contribute to improved operational intelligence and data analysis results. Manufacturing data quality also ensures the needed accuracy and reliability of the data measured along the value chain. Data quality in manufacturing boosts (i) product quality in the supply chain; and (ii) process quality of the manufacturing companies. Data quality in i4Q includes systematically identifying the factors that influence data quality in manufacturing by using data quality management and technologies supporting it.

This deliverable contains a **guideline to manage data quality in manufacturing**. It introduces the conceptual basis, including data and information definitions, as well as related quality and management concepts. Besides, it outlines an activity framework to plan an information flow analysis that results in a collection of data quality factors to manage. This deliverable identifies three activity types that provide means to influence quality factors and maintain high data quality.

The **target audience** for this document includes quality managers, researchers, and production managers.

1.2 Goals

The following list summarizes this deliverable's primary goals:

1. Provide key activities to manage data quality in manufacturing
2. Create a methodological connection to other tasks (mainly in WP3)

Goal 1 concerns the external benefits of this document. It focuses on delivering a set of easy-to-understand activities that various end-users could apply. These activities must address resource constraints in organizations – especially smaller producers with limited data quality management budgets. Simultaneously, the key activities should ground on acknowledged theories.

Goal 2 focuses on the internal effects of this document. It means organizing the activities in WP3 (and some other tasks in WP4 and WP5) along with a common framework. The scope includes activities that rely on software and those that do not.

2. Conceptual Basis

2.1 Background

This guideline uses the **quality concept** outlined in the ISO 9000 standard series (ISO 9000:2015, 2015). It assumes that quality is the degree (or match) between a thing's actual characteristics and stated requirements for these. For instance, a thing can be a product with a defined geometry (required characteristic). In this simple case, the product's quality is the deviation between its actual (as produced) and required geometry. The match can be gradual, such as 95%, or binary like fulfilled and not fulfilled. **Characteristics** are features capable of distinguishing one thing from another - not all features are characteristics.

Data and information management literature adapted the notion above and transferred it to data and information. At some point, the International Organization for Standardization (ISO) started developing and publishing standards related to data quality. Their standards are broadly recognized and applied and, therefore, a good grounding for this guideline. Two standard series focused on data quality ground this document:

- **ISO/IEC 25012** is a standard focusing on structured data's quality (ISO/IEC 25012:2008, 2008). It covers all data types, assigned data values, and relationships between data. The standard excludes short-lived (not persisted) data from embedded devices or real-time sensors. Furthermore, it excludes the metadata that ISO/IEC 11179 (ISO/IEC 11179-1:2015, 2015) covers. Besides, it focuses on data as part of a computer system. ISO/IEC 25012 belongs to the 25000 standard series dedicated to "*Software product Quality Requirements and Evaluation (SQuaRE)*".
- **ISO 8000** is a standard for master data quality management (ISO 8000-1:2022, 2022). *Master data* concerns the fundamental facts about an organization's customers, products, employees, suppliers, services, shareholders, facilities, equipment, and rules and regulations. ISO 8000 aims to extend and clarify ISO 9001 (ISO 9001:2015, 2015) and excludes software product quality (ISO/IEC 25000 series). The ISO 8000 standard series introduces terms and definitions different from ISO 9000:2015.

This document adopts the definitions of ISO 9000 and ISO/IEC 25012, which are more in line with information science. However, ISO 8000 contributes with conceptual extensions and certain clarifications helpful for this document's context.

The remainder of this section introduces the key terms and concepts used in this document. Deliverable 5.6 (i4Q Manufacturing Line Data Certification Procedure) uses, adapts, and extends them.

2.1.1 Data and information

There is no generally acknowledged definition for data or information. Definitions differ among the disciplines, and practitioners often use both terms synonymously. This guideline uses definitions closer to the stricter ones proposed in information science for two reasons.

- First, this guideline also outlines technical measures to maintain high data quality. Explaining them requires technical depth that benefits from a stricter definition of "data".

- Second, this guideline focuses on the application of information as it aims to improve product and process quality. The stricter differentiation of data and information allows clearer system boundaries and assigning roles and activities.

Definitions of data and information often include or imply a hierarchy, as illustrated in Figure 1 (Frické 2009).[1] This hierarchy may include other concepts, such as knowledge or wisdom, indicated by the ellipsis at the pyramid's tip.

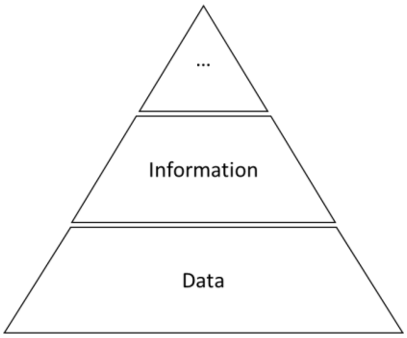


Figure 1. Data, information, and "something else" pyramid

In this guideline, the critical aspect is that information requires data – often, one information item consists of many data items. The following definitions originate from the ISO 9000 standard "Quality management systems – fundamentals and vocabulary". Table 1 summarizes the definitions and provides examples.

Terms	Definitions	Examples
Data	"Facts about an object" (ISO 9000:2015, 2015)	"100° Celsius oil temperature"; "Cutting oil in workstation A"
Information	"Meaningful data" (ISO 9000:2015, 2015)	In a dashboard: "Workstation A's cutting oil is 100° Celsius"

Table 1. Data and information definitions and examples

An advantage of the definitions above is that information grounds on data, and, therefore, the guideline can use the term "data" in a broader sense. This decision increases this guideline's readability – the remainder of this document will refer to data unless differentiation is helpful. The example above demonstrates how a dashboard could use two facts about objects to create meaning. Of course, further data is necessary for most application areas. Besides, dissecting an example typically involves other data stakeholders, and this multi-perspective process often leads to a compromise that works for the specific organization.

In the context of ISO/IEC 25012, data and information exist in a computer. Many data items will exist outside of a computer, at least for some time. Examples include printed forms, instructions, manuals, and notes.

[1] This hierarchy leaves out symbols as the constituents of data.

This guideline covers the **data inside and outside computers** to have broad relevance.

2.1.2 Data life cycle

One of the key concepts in the ISO 9000 series is the **process**, i.e., a set of interrelated or interacting activities. Controlling processes is essential to systematically creating the results the organization needs. In information management, **life cycle models** organize the processes from data creation to destruction (similar to life cycles in biology). These models serve different purposes and, therefore, are not generally acknowledged.

ISO 25024:2015 introduces a data life cycle beginning with data design and ending with data deletion (ISO 25024:2015). This standard excludes knowledge representation, data mining techniques, and statistical significance for a random sample, but its concepts are highly relevant to this guideline. Figure 2 illustrates data-related processes and their relations as a data life cycle.

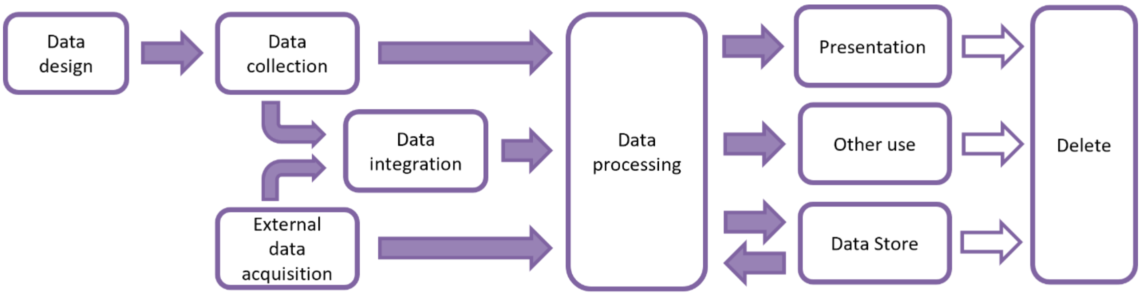


Figure 2. Example of a data life cycle (ISO 25024:2015)

This guideline adapts the life cycle model presented in ISO 25024:2015. This adapted model assumes that all data is a construct from a human or a technical system, i.e., a machine. Technical systems include measurement systems and computers measuring through software. Humans design, build, deploy, and maintain these systems – they are responsible for their results. Figure 3 summarizes the assumption above in a conceptual framework.

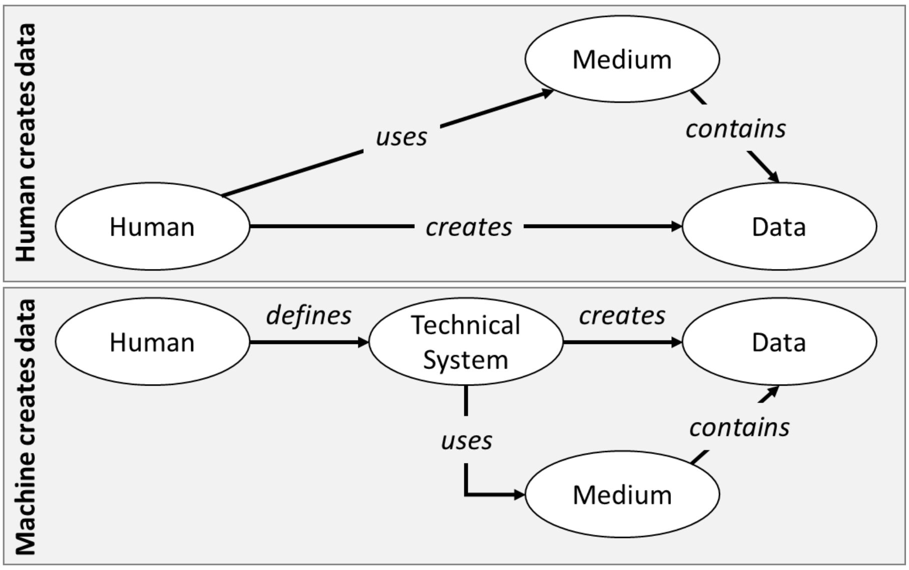


Figure 3. Humans and machines creating data

As data and information cannot exist alone, they need a medium that contains or transports them. Media can be software too, and humans will need interfaces to create data. In this case, they use human-computer interfaces (HCI) with specific characteristics to formulate the data (e.g., filling a form, writing a program, and drawing). Humans, HCI, the technical system, and the medium affect the features of data items. Other processes along the data life cycle further influence these features. The following list summarizes critical processes:

- **Data storage** preserves data so that users can use it later
- **Data processing** covers systematic operations upon data. It includes arithmetic or logic operations, merging or sorting, assembling or compiling of programs, or operations on text, such as editing, sorting, merging, storing, retrieving, displaying, or printing (ISO/IEC 2382:2015, 2015)
- **Data integration** means combining data from heterogeneous sources and providing a unified view of them (Lenzerini 2002)
- **Data deletion** is the final process in the life cycle that destroys data permanently

I4Q has several solutions that rely on **machine learning** (ML). Therefore, this guideline emphasizes it. ML is an approach that lets software programs learn and is often one step toward realizing complex data processing. Besides, it is essential for building *artificial intelligence* (AI). With ML, an organization can train software to analyze images, time series, natural language, or predict events and states. There is supervised and unsupervised ML.[2]

Machine learning

"Process by which a functional unit improves its performance by acquiring new knowledge or skills, or by reorganizing existing knowledge or skills" (ISO/IEC 2382:2015, 2015)

Supervised machine learning relies on correct training data that a computer uses to learn patterns present in this data (Braga-Neto 2020). It results in a so-called model that computers can apply to new data to, for instance, classify that data or predict something. Figure 4 illustrates the design cycle for supervised machine learning.

[2] This document does not cover reinforced ML to reduce complexity and increase understandability.

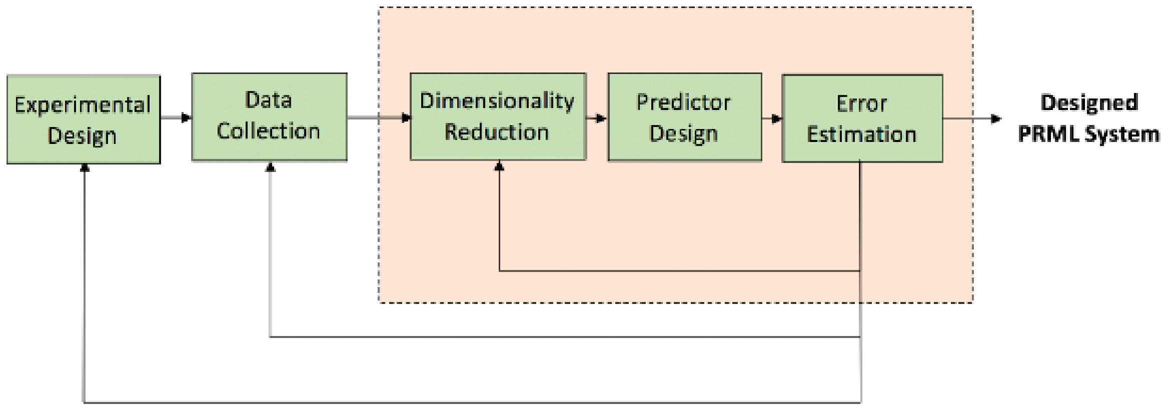


Figure 4. Design cycle for supervised machine learning (Braga-Neto 2020)

Designing for supervised ML assumes an experimental design before the experimenter collects the data to train the model. This step includes framing the question, identifying the populations and relevant features, and determining the appropriate sample sizes and sampling mechanisms. A supervised ML process' results contain *different kinds of error*s, such as random and expected errors. Managing these errors is an essential skill of a data scientist.

Unsupervised machine learning means there is no data labeled as correct. Its main purpose is to detect structure in data – the structure can be so complex that humans cannot identify it. Measuring performance is much more challenging in this machine learning case.

Figure 5 concludes the data life cycle used in this guideline. It integrates the aspect of machine learning to account for its relevance in companies using AI.

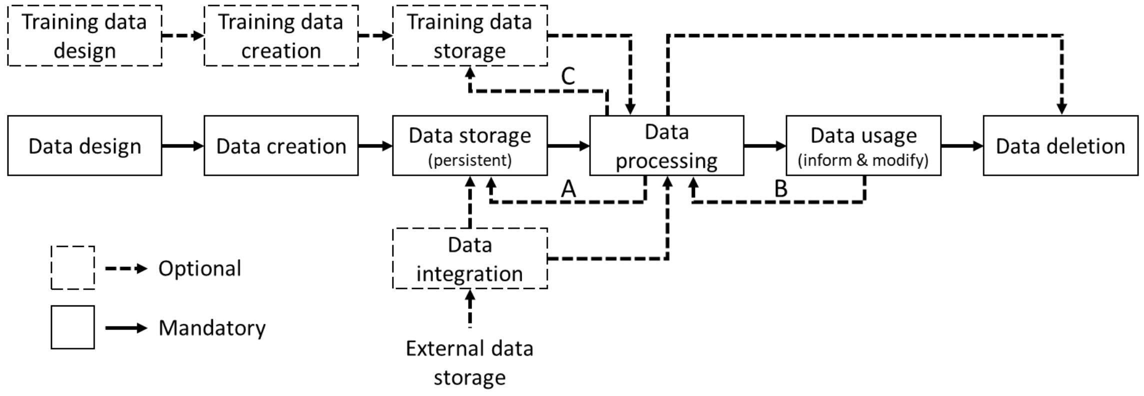


Figure 5. Data life cycle model with machine learning and data integration

The life cycle model above includes *mandatory* and *optional* elements and relations. Optional processes refer to ML and data integration. They add flexibility and specificity to the model.

Data and training data design are planning steps and refer to the situation when there is no data item. The subsequent steps refer to data items. **Data usage** refers to situations where data *informs* a user and where a user *modifies* data. **Data deletion** means the destruction of the data item either triggered by modification or data processing.

Backward-directed arrows (A, B, and C) indicate loops where data returns to the preceding process. **Arrow A** refers to situations where, for instance, an organization does not use certain processed data but stores it for later (or without ever using it). **Arrow B** indicates updating data, which could trigger arrow A to store the updated data. **Arrow C** outlines the situation where data processing affects training data. This loop can describe cases where an AI adapts its training data based on data processing (e.g., removing or adding labeled examples).

Data processing is a single element in the model above. In practice, other steps typically include processing data for technical reasons, e.g., processing raw measurements and filtering or cleaning data. The life cycle model above does not include this form of data processing to reduce the model's complexity. Likewise, **Data distribution** is not explicitly covered. Distribution allows the software to receive or access the stored or processed data.

2.1.3 Information needs and users

This section clarifies data usage and how it emerges. The starting point is the information need concept. An **information need** is "*a hypothesized state brought about when an individual realizes that they are not comfortable with their current state of knowledge*" (Case and Given 2016; Taylor 1968). This need exists in different forms, as illustrated in Figure 6.

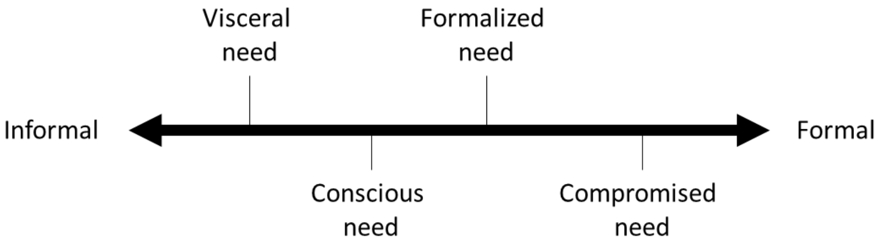


Figure 6. Information need typology (based on Taylor 1968)

A visceral need is a vague feeling, while the conscious need describes a mental description of the information needed. After refinement, this need evolves into a formalized need that an information user can express. Fulfilling the need will involve working with software that may not meet all aspects of the formal need. For instance, the data provided by the software is not that precise but still usable due to a lack of alternatives – the formalized need turns into a compromised need.

This guideline assumes that information users can formally describe their information needs.

2.1.4 Data and information quality

A key challenge of applying the ISO 25012 quality concept to data and information is the **selection of suitable characteristics**. Various related articles and books contain at least one preferred list of characteristics (i.e., data quality model) - there is no generally accepted list. This guideline uses the characteristics described in ISO/IEC 25012:2008. **Table 2** summarizes the characteristics, descriptions, and views on them. Descriptions typically follow this pattern: "The degree to which <*specific part*> in a specific context of use.". This guideline simplifies the descriptions' sentence structures without changing their meanings.

Characteristics	Description	Views	
	The degree to which data <specific part> in a specific context of use.	I	SD
Accuracy	[...] data has attributes that correctly represent the true value of the intended attributes of a concept or event [...]	X	
Completeness	[...] subject data associated with an entity has values for all expected attributes and related entity instances [...]	X	
Consistency	[...] data has attributes free from contradiction and coherent with other data [...]	X	
Credibility	[...] data has attributes that users regard as true and believable [...]	X	
Currentness	[...] data has attributes of the right age [...]	X	
Accessibility	[...] data can be accessed [...], particularly by people who need supporting technology or special configuration because of some disability.	X	X
Compliance	[...] data has attributes that adhere to standards, conventions, or regulations in force, and similar rules relating to data quality [...]	X	X
Confidentiality[3]	[...] data has attributes that ensure that it is only accessible and interpretable by authorized users [...]	X	X
Efficiency	[...] data has attributes that can be processed and provide the expected performance levels by using the appropriate amounts and types of resources [...]	X	X
Precision	[...] data has exact attributes or that provide discrimination [...]	X	X
Traceability	[...] data has attributes that provide an audit trail of access to the data and of any changes made to the data [...]	X	X
Understandability[4]	[...] data has attributes that enable users to read and interpret it, and are expressed in appropriate languages, symbols, and units [...]	X	X
Availability	[...] data has attributes that enable authorized users and applications to retrieve it [...]		X

[3] Confidentiality is an aspect of information security and therefore connected to Task 3.4.

[4] Understandability sometimes depends on metadata.

Characteristics	Description	Views	
	The degree to which data <specific part> in a specific context of use.	I	SD
Portability	[...] data has attributes that enable it to be installed, replaced or moved from one system to another, preserving the existing quality [...]		X
Recoverability	[...] data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, [...]		X

Table 2. Data quality model characteristics (ISO 25024:2015). I: Inherent; SD: System dependent

The views on characteristics are relevant because they influence which activities effectively change a characteristic. **Inherent** characteristics deliver some benefit in themselves when someone or something uses data. This view does not require a computer to store data - they could be on paper, for instance, and would still be helpful. The **system dependent** view covers how hardware and software affect data characteristics. For instance, availability entirely depends on the system preserving and providing data.

The importance of characteristics differs among stakeholders and steps in the data life cycle. A helpful concept to clarify this aspect is the so-called **Evolutional Data Quality** (EDQ) concept (Liu and Chi 2002).

Liu and Chi (2002) developed a theory-based view on data quality that focuses on the evolution of data along a life cycle. Their data evolution life cycle contains four phases:

- **Data collection** concerns data capturing through observation of real-world processes, measurement, and perception.
- **Data organization** means structuring and storing data in files, databases, and other forms of data storage.
- **Data presentation** subsumes processing, interpretation, summarizing, formatting, and presentation of data in views.
- **Data application** is the final phase where users utilize data, which can trigger further data collection.

Their concept assumes that quality characteristics relevant to a phase contribute to the characteristics of all following phases. This assumption implies a cause-effect diagram (i.e., an acknowledged quality management instrument) like **Figure 7**. The cause-effect diagram has the structure of a data life cycle – such diagrams can use different life cycle models.

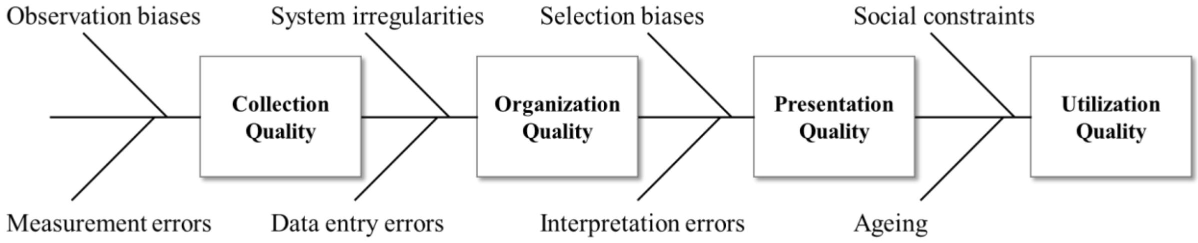


Figure 7. Cause-effect chain in the Evolutional Data Quality concept (Liu and Chi 2002)

The notion above implies that causes affect data characteristics and related data quality. Some quality changes may emerge when stakeholders change requirements for data. **Figure 8** illustrates an extended EDQ concept with causes affecting characteristics and requirements to account for this specific situation.

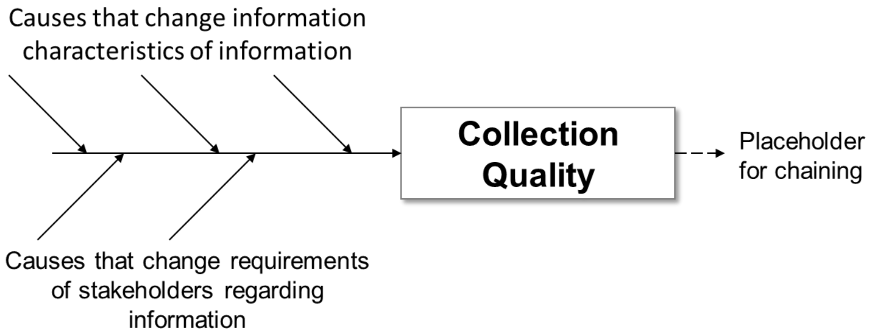

Causes that change information characteristics of information
Collection Quality
Placeholder for chaining
Causes that change requirements of stakeholders regarding information

Figure 8. Extended Evolutional Data Quality concept

This deliverable and D3.2 "i4Q QualiExplore for Data Quality Factor Knowledge" refer to the causes above as "**quality factors**". An organization must manage these factors to influence data and information quality. This document focuses on causes that change information characteristics.

An organization must manage the factors influencing information quality characteristics.

2.1.5 Data and information quality management

Data and information quality management are not the same. The main difference is that **data quality management** (DQM) focuses on technical aspects of storing and organizing data, while **information quality management** (IQM) primarily concerns information application.

IQM belongs to the organization's information management (IM) process. IM manages an organization's processes, resources, technologies, and policies, focusing on information (Choo 2000). It prepares, realizes, and monitors information systems that provide information to employees and stakeholders. The concept is much broader in comparison with DQM. IQM promotes a user-centered view and emphasizes the understandability and usability of the information. The broad scope of IM means that IQM must **consider various factors** influencing information quality. They include:

- collection, organization, distribution, and application of information (processes)
- employee behavior and the available IT infrastructures (resources)
- advantages and disadvantages of data processing methods (technologies)
- security and privacy regulations and governance models (policies)

These factors refer to the same "quality factors" concept outlined in the EDQ. IQM's broader scope and focus on processes and information items make it the right approach to managing the quality of manufacturing data in i4Q.

A closely related management concept is **corporate data/information governance**. Generally, corporate governance deals with an organization's rules, practices, and processes (Chen 2021). It balances stakeholder interests and, due to its broad scope, affects all areas of a company. Different models detail corporate governance focused on data and information (Krcmar 2015). A common characteristic is that this governance branch defines the roles and organizational structures to make the most out of an organization's information. Table 3 outlines relevant roles for i4Q based on the descriptions provided in the ISO 8000 standard series.

Roles	Descriptions (based on ISO 8000-2 and ISO 8000-150)
Data manager	Directs a data quality management plan aligned with the organization's objectives, regulates factors affecting data quality at the organizational level, and defines plans for data quality processes and support activities • Grants data administrators authority to trace and correct data • Analyses factors affecting data quality in data planning • Improves business processes
Data administrator	Defines guidelines for maintaining data quality and avoiding recurrence of data errors by evaluating reasons for errors, eliminating root causes, or developing data schema • Controls and coordinates data technicians • Conducts root cause analysis to identify data quality issues • Carries out data quality management plan
Data technician	Creates, reads, updates, and deletes data following the data administrator's data quality management procedures • Measures data quality • Corrects incorrect data • Prepares reports, retrieves required information, and deletes outdated data

Table 3. Roles related to data quality management

2.1.6 Production and production data

Production is the process of transforming inputs into outputs. Producers are organizations producing goods from raw materials and semi-finished products. They define the production parameters to ensure the final products meet or exceed requirements, i.e., have high quality. Figure 9 illustrates system levels for a producer, from the broad scope of a production network to individual workstations and production machines.

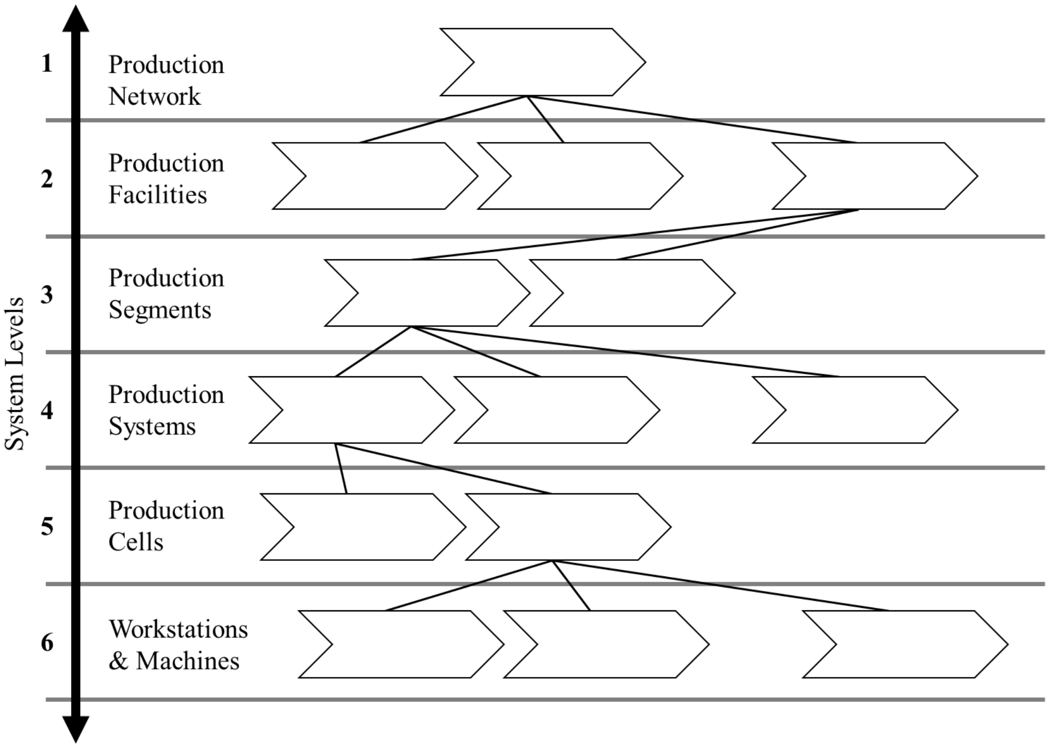


Figure 9. System levels of a producer (Westkämper 2008)

Data quality concerns all levels, but this guideline's scope is on the production system and levels below it. This scope is most helpful in balancing the complexity and coverage of different data. A production system has system layers, as shown in Figure 10.

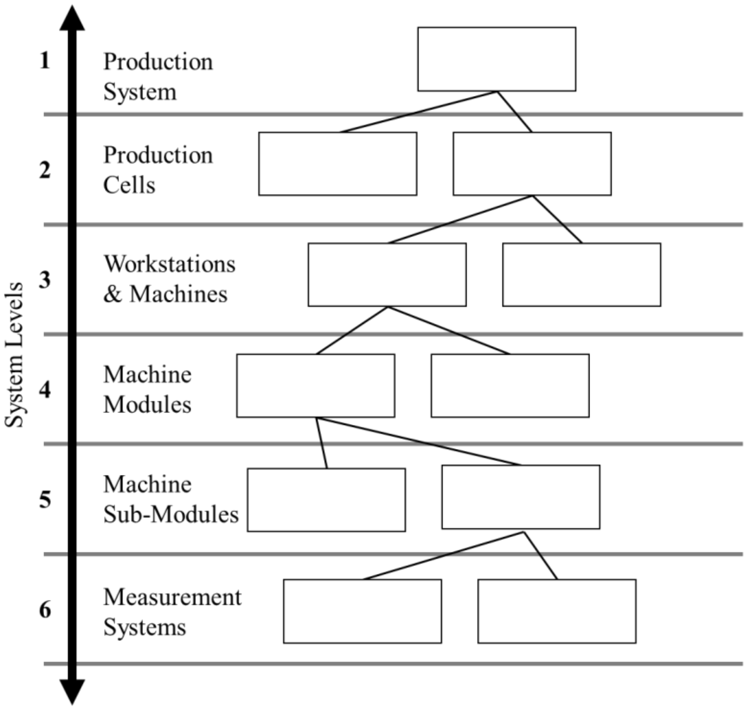


Figure 10. System levels in a production system (Nyhuis 2008; Westkämper 2008)

The lowest level in this production system model represents measurement systems (i.e., sensors). They are the source of information about, for instance, processes, events, system and environment states, and locations. Humans are not visible in the illustrations above but perform tasks on various layers, including creating information, for instance, by filling out forms.

This guideline focuses on production systems and lower system layers.

2.2 Basic guideline structure, stakeholders, and roles

This section introduces the basic structure of the management process that grounds this document. Besides, it re-introduces the stakeholders identified in WP2 and combines them with typical roles in IQM.

2.2.1 Plan-Do-Study-Act (PDSA)

The PDSA cycle[5] is the basic management process behind the ISO 9001:2015 standard (N.a. 2022). Figure 11 illustrates this cycle, and the following paragraphs outline each step.

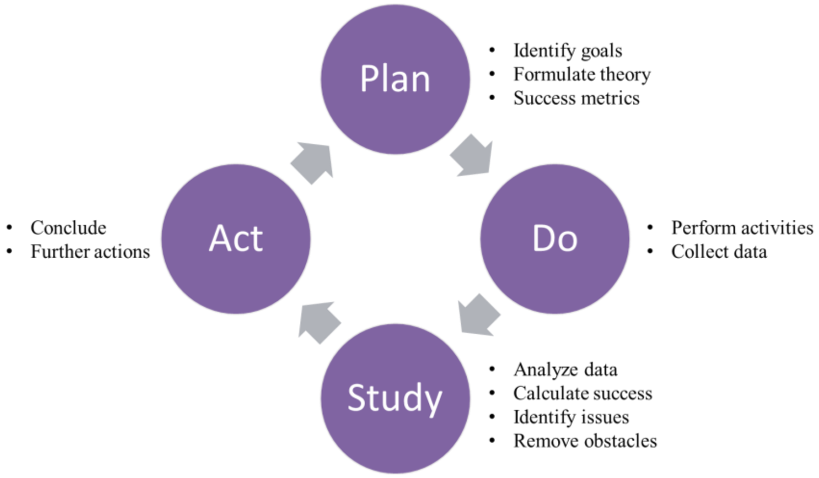


Figure 11. The PDSA cycle

Plan. The first step in this cycle identifies goals, formulates a theory, and defines success metrics for information quality. It also plans the activities to realize the goals, such as new or revised functions and organizational procedures, and the collection of data needed to assess the progress against the goals.

- Goals clarify how the data stakeholders want the information to be. Reaching the goal means achieving a change in information quality (improvement).
- The theory outlines, for instance, how quality problems emerge and relate.
- Success metrics specify numbers under which conditions the information quality fulfills the goals.

[5] PDSA is Deming's updated version of the PDCA cycle. He replaced the Check step with the Study step. Checking implies verification of a plan rather than learning from failure.

Do. In this step, the data quality expert performs the planned activities to improve information quality. This activity includes software changes (e.g., improved user interfaces) and non-functional procedures (e.g., user training). The staff also collects the data needed to assess the success metrics.

Study. The third step analyses the collected data and calculates the success metrics. It identifies issues in the plan and removes obstacles that hamper achieving the goals. Potential issues include human and computational resource bottlenecks and interference from production system changes.

Act. This step concludes the study results and identifies further actions to reach the goals. It can also change goals. A new cycle starts with new goals or adapted ones. In the light of cybersecurity, this step includes a variety of actions to prevent and detect attacks that affect data integrity. They range from implementing audit trails to establishing management security qualifications and maintenance programs.

This guideline focuses on the Plan and Do steps.
Plan refers to an information flow analysis (Section 3.1), and **Do** (Section 3.2) concerns measures to manage data quality factors.

2.2.2 Stakeholders and Roles

This section focuses on stakeholders, i.e., parties interested in production data quality, and the assignment of data quality management roles. D2.3 already identified and described the primary stakeholders in i4Q. Figure 12 summarizes these stakeholders' relations as a reminder.

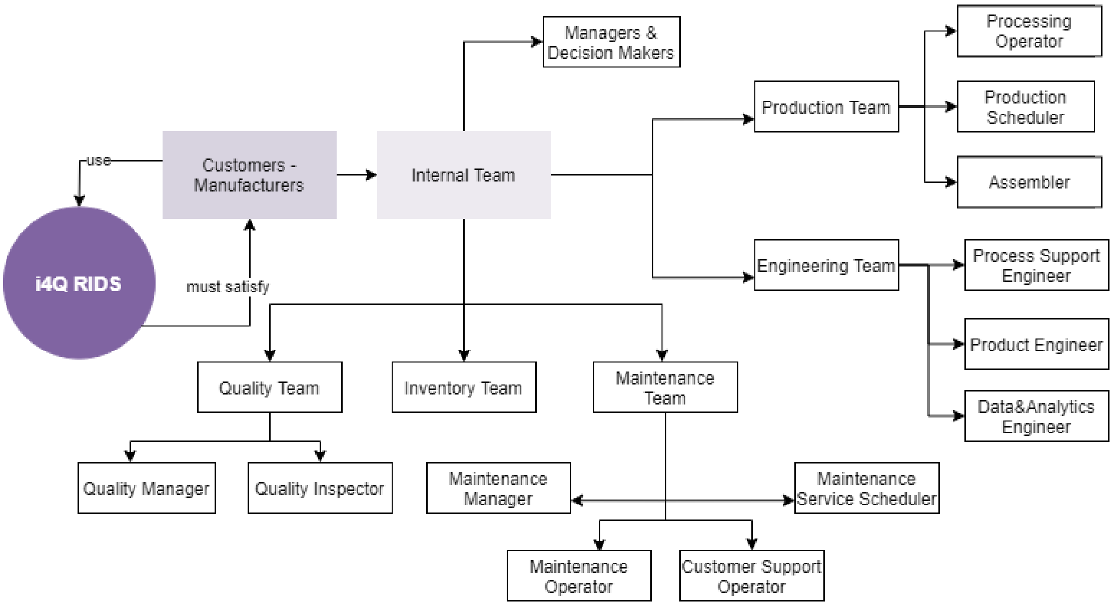


Figure 12. Relation of primary stakeholders in i4Q (from D2.3)

D2.3 also outlined the primary stakeholders' "wants" from the i4Q solutions. Some of them refer directly to information needs or data creation activities. The following paragraphs use the terms "data producer" and "data consumer" to reflect these general roles. In general, there will be significantly more consumers than producers – which accounts for information sharing among the different stakeholders in an organization.

Table 4 summarizes the stakeholder's wants above and general data roles. Since this table focuses on the i4Q solutions' wants, it does not mention general information needs. Nevertheless, such information is essential for production in general. Guideline users should adjust the wants as needed to best reflect their organization's needs – this deliverable's second version will update the table below. **Table 3**We will collect this information in collaboration with the end users during test rounds (e.g., workshop or interview).

Stakeholder Names	Functional Capabilities I want to:	Produces data?	Consumes data?
Process Support Engineer	Identify factors that influence the quality	N	Y
	Predict possible product problems	N	Y
Processing Operator	Be notified when deviations from standard functioning values occur	N	Y
	Simply modify process input configurations	Y	N
Production Scheduler	Receive information on the production capacity and resource availability	N	Y
	Have support and suggestions for the production schedule definition	N	N
	Receive feedback from actual production	N	Y
	Receive feedback on the quality of the final product	N	Y
	Have support for the production schedule update	N	N
Assembler	Have support to test the output to ensure the highest quality	N	N
	Receive feedback and suggestions for improving the quality of the output	N	Y
	Report on issues, malfunction or defective parts	Y	N
Product Engineer	Identify factors that influence the quality and/or functionality of a product	N	Y
	Evaluate the new/updated product in terms of functionality and quality	N	N
	Visualise and compare performance, reliability and costs of materials and/or suppliers	N	Y
	Have support to determine production costs of the new/improved product	N	N
Data & Analytics Engineer	Develop high performance data pipelines to support complex data integration	Y	N
	Oversee ETL (extract, transform, load)	N	N
	Build and train data models	Y	N

Stakeholder Names	Functional Capabilities I want to:	Produces data?	Consumes data?
	Analyze multiple data sources in detail to identify quality trends and problem indicators	N	Y
	Receive suggestions for processes improvement	N	Y
Quality Manager	Certify the quality of the process in a simple and verifiable way	N	N
	Certify product quality in a simple and verifiable way	N	N
	Visualize information about the quality of item or process	N	Y
	Identify the potential origin of an issue in a simple way	N	N
	Have support for the final decision on a quality issue	N	N
Quality Inspector	Visualise information about an item or process	N	Y
	Perform the testing of incoming raw material in a simple but accurate way	N	N
	Perform testing of a product in a simple but accurate way	N	N
	Report and save the result of the evaluation	Y	N
	Have support on decision concerning escalation	N	N
Maintenance Manager	Forecast the maintenance expenditure and prepare a budget to ensure that maintenance expenditure is as per planned budget	N	Y
	Receive information and suggestions regarding the maintenance activities	N	Y
Maintenance Service Scheduler	Receive suggestions to schedule the maintenance work (after due consultation with the concerned production departments)	N	Y
	Prepare an inventory list of spare parts and materials required for maintenance	Y	N
	Ensure proper inventory control of spare parts and other materials required	N	N
	Monitor the equipment condition at regular intervals	N	Y
Maintenance Operator	Receive information and support to carry out repairs	N	Y
	Provide feedback concerning the maintenance suggestions	Y	N

Stakeholder Names	Functional Capabilities I want to:	Produces data?	Consumes data?
	Be notified of the acquisition, installation and operation of machinery	N	Y
	Document and maintain a record of each maintenance activity (i.e., repairs, replacement, overhauls, modifications and lubrication etc.)	Y	N
Customer support operator	Manage customer reports (ticketing system)	Y	N
	Receive information and support to analyse the problem	N	Y
	Have support to decide whether to implement maintenance procedures	N	N
Inventory Team	Examine the levels of supplies, raw material and final products to determine shortages	N	Y
	Receive feedback on the quality of raw material	N	Y
	Visualise and compare performance, reliability and costs of materials and/or suppliers	N	Y
	Receive support for preparing the notification of the quality of the material to the supplier	N	N
	Receive information to prepare detailed reports on inventory operations, stock levels, and adjustments	N	Y
	Perform daily analysis to predict potential inventory problems	N	Y

Table 4. Stakeholders and their wants and general data roles in i4Q

Besides the general data roles above, stakeholders can have one or more roles related to data quality management. **Table 3**Section 2.1.5 outlined the relevant role descriptions already. Table 5 combines these roles with the stakeholders above to create a template that will indicate relevant assignments for i4Q. Ideally, a production has stakeholders that exercise all data-related roles at least once (likely for specific areas in a production). The second version of this deliverable will contain an update of this table with all relevant stakeholders and relevant role assignments. We will collect this information in collaboration with the end users during test rounds (e.g., workshop or interview).

	Data manager	Data administrator	Data technician
Processing Operator			
Assembler			
Data & Analytics Engineer			

	Data manager	Data administrator	Data technician
Quality Inspector			
Maintenance Service Scheduler			
Maintenance Operator			
…			

Table 5. Stakeholders and relevant data quality management roles in i4Q (template, no assignments yet)

3. Activity Framework

This section outlines the suggestions for an information flow analysis and proposes types of activities to maintain data quality. Section 3.1 matches the Plan step in the PDSA cycle and results in success metrics for data quality and quality factors to control. Section 3.2 relates to the Do step and covers activities influencing data quality factors.

3.1 Information flow analysis

The information flow analysis has several essential steps outlined in the following paragraphs.

The **first** step is setting the **analysis scope** by selecting the target stakeholders and the production system boundaries. Table 6 illustrates an example table to describe the scope of the information flow analysis. Organizations can customize it to their needs and may also decide to focus on the lowest level without any specific stakeholder in mind.

	Produces data?	Consumes data?	Production systems	Production cells	Workstations	Workstation modules	Workstation sub-modules	Measurement systems
Process Support Engineer	N	Y		X				
Processing Operator	Y	Y			X			
Production Scheduler	N	Y		X				
Assembler	Y	Y			X			
Product Engineer	N	Y						
Data & Analytics Engineer	Y	Y			X			
Quality Manager	N	Y		X				
Quality Inspector	Y	Y		X				
Maintenance Manager	N	Y						
Maintenance Service Scheduler	Y	Y						
Maintenance Operator	Y	Y						
Customer support operator	N	Y						
Inventory Team	N	Y						
No specific stakeholder	-	-						X

Table 6. Primary stakeholders and production system level (example analysis scope)

The **second** step is **identifying the formalized information needs** of the target stakeholders within the production system scope. This step uses interviews, questionnaires, or document analysis. The resulting need descriptions should contain data quality characteristics to provide success metrics for data quality. These metrics should refer to the data quality characteristics outlined in Section 2.1.4.

The **third** step focuses on the data life cycle processes and outlines **the system boundaries** from this perspective. Figure 13 summarizes the processes introduced in Section 2.1.2. Some processes are mandatory, and others could be relevant.

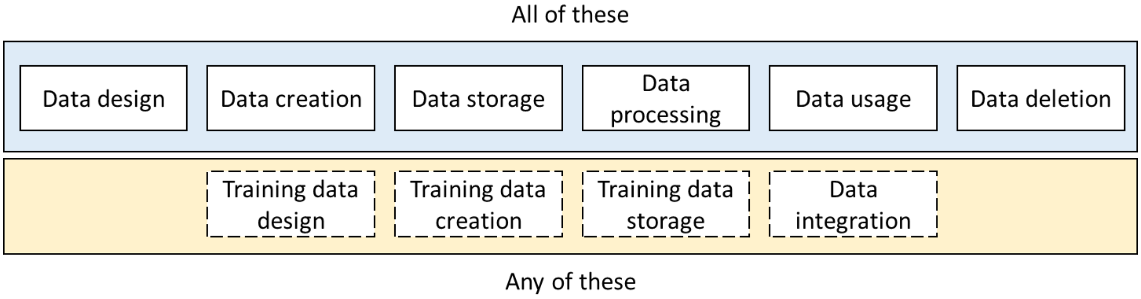


Figure 13. System boundary definition

The **fourth** step **investigates each selected process**. This analysis focuses on identifying relevant quality factors that roles within the organization can influence. Finding quality factors is challenging since there are so many, and often they are not obvious, such as the many biases potentially introduced to datasets. Table 7 summarizes aspects to consider during the information flow analysis to guide the investigation. Collections of data quality factors, such as the one in QualiExplore (D3.2), can further support this process.

	Design*	Creation*	Storage*	Processing	Usage	Deletion	Integration
Dependencies with third parties			X	X		X	X
Human-in-the-loop activities	X	X	X			X	
Application of machine learning				X			
* *Focus on common production data and training data for machine learning*							

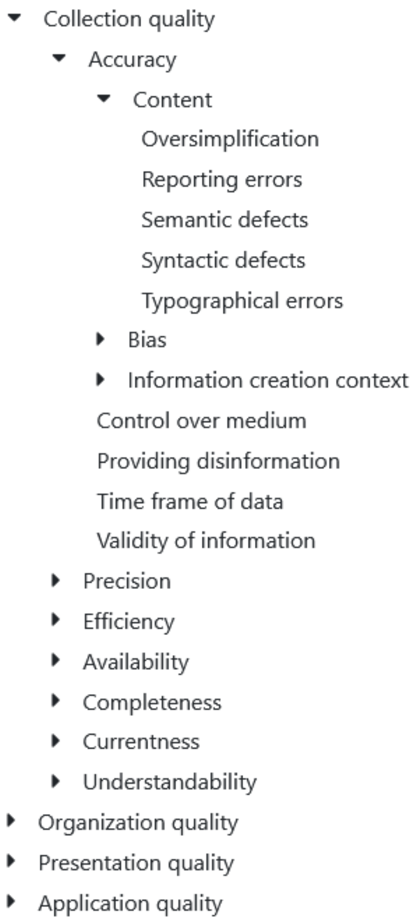
Table 7. Aspects to consider during information flow analysis

Dependencies with third parties affect, for instance, the accessibility and timeliness of the information. Services may become temporarily unavailable or slower than agreed. Service level agreements typically address these risks. Other issues may occur when third parties create information for the producer, e.g., a customer service sub-contractor reporting issues to the producer. In this situation, the producer does not directly influence characteristics such as completeness, accuracy, and timeliness of reported information. Measures to minimize information quality problems may be more complex (e.g., specific sub-contractor selection criteria, report review processes).

Human-in-the-loop activities refer to control loops where humans must decide before the data-related process proceeds. These decisions relate to factors of human behavior that may introduce bias or errors in datasets. Therefore, it is essential to investigate human involvement thoroughly. Related life cycle processes are a) data design when humans define scopes and relations, b) data creation when humans decide, for instance, which information they report, c) data storage where humans decide which documents they upload, and d) data deletion when humans decide which data they delete.

ML has become more frequent among producers, including producers using software relying on ML and on producers who exercise ML themselves. In the first case, the producer depends on a third party. Biases are the most critical and controversial factors related to ML. Producers must be cautious about them, especially if the organization does not have substantial experience with ML yet.

The information flow analysis result is a collection of quality factors per investigated life cycle process. One technique to illustrate these factors is a tree where branches represent EDQ concepts, quality characteristics, and factor groups. Its leaves are individual factors. **Figure 14** illustrates an example tree. Producers must assess each factor and decide if they want to manage it by implementing quality increasing or maintaining measures.



- ▾ Collection quality
  - ▾ Accuracy
    - ▾ Content
      - Oversimplification
      - Reporting errors
      - Semantic defects
      - Syntactic defects
      - Typographical errors
    - ▸ Bias
    - ▸ Information creation context
    - Control over medium
    - Providing disinformation
    - Time frame of data
    - Validity of information
  - ▸ Precision
  - ▸ Efficiency
  - ▸ Availability
  - ▸ Completeness
  - ▸ Currentness
  - ▸ Understandability
- ▸ Organization quality
- ▸ Presentation quality
- ▸ Application quality

Figure 14. Example tree with EDQ, quality characteristics, factors groups, and factors

3.2 Measures to influence quality factors

Once the producer knows the relevant quality factors, its employees can identify measures to influence them. The goal is to find cost-effective measures that increase data quality.

Section 2.1.2 indicated that humans are directly or indirectly involved in creating data. Similarly, they influence other life cycle processes where they design and operate information systems. Consequently, the root cause of information quality problems is humans, and mitigation measures should aim to change human behavior.

A producer has **various instruments to influence its employees' behavior** regarding data quality. They include, for instance, formal company guidelines and informal rules and acknowledged practices. For example, a quality management Wiki (handbook) should contain knowledge on how employees must exercise specific tasks. In terms of software, producers use features to ensure that employees provide the right information. Common features are input validation and auto-complete to ensure correct and consistent information. Finally, a producer also manages the **behavior of third parties** via contracts and an organization's guidelines. The latter is relevant, for instance, when sub-contractors work on-site, such as maintenance service providers. Some third parties, such as policymakers, are not subject to contracts, and the producer may have minimal influence on their behavior.[6] The remainder of this document focuses on the producer's employees.

Figure 15 summarizes this guidelines activity framework. It adapts the framework developed for the Horizon 2020 research and innovation action NIMBLE (NIMBLE Consortium 2020).

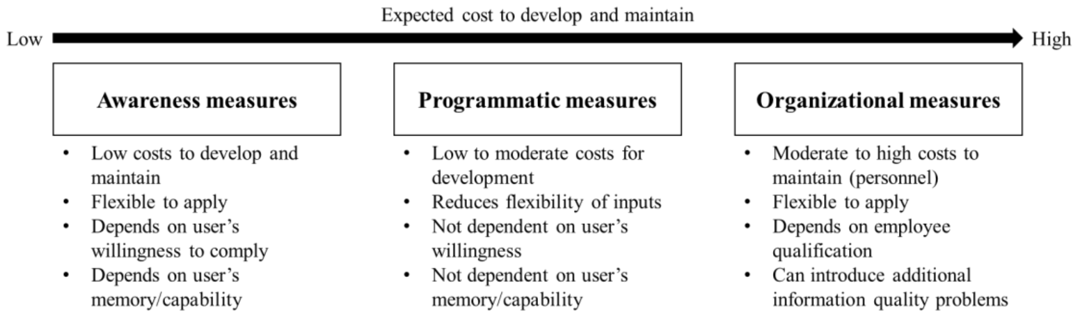

Expected cost to develop and maintain
Low
High
Awareness measures
• Low costs to develop and maintain
• Flexible to apply
• Depends on user's willingness to comply
• Depends on user's memory/capability
Programmatic measures
• Low to moderate costs for development
• Reduces flexibility of inputs
• Not dependent on user's willingness
• Not dependent on user's memory/capability
Organizational measures
• Moderate to high costs to maintain (personnel)
• Flexible to apply
• Depends on employee qualification
• Can introduce additional information quality problems

Figure 15. Activity Framework

The framework has three activity groups organized along a continuum of the expected cost to develop and maintain measures. These groups represent the producer's core instruments influencing its employees and, consequently, relevant data quality factors.

The following sub-sections outline the activity groups above and the relevant factors they could influence. Descriptions do not suggest specific measures, and several measures can affect the same factor. A comprehensive summary is neither intended nor feasible for this guideline. This deliverable's second version will update the measures based on end-user feedback.

3.2.1 Awareness measures

These measures are cheap to develop and maintain because they do not require deep integration in software – e.g., static websites with information could be sufficient to raise awareness. Awareness measures are flexible because one solution can make users aware of various topics. The downside of this measure is that it depends on each user's willingness and capability to behave in a way that minimizes data quality problems. Consequently, this characteristic makes these measures less reliable unless the measures are recurring and regularly controlled. Table 8 outlines example factors for this type of measure. The second version of this deliverable will update the table.

Quality Characteristics	Quality Factors	Description
Accuracy	Sample bias	The sampling process produced a dataset that misrepresents the target population's characteristics.
Accessibility	Willingness to share information	Information authors or providers must see value in sharing information.

Table 8. Example factors influenced by awareness measures

[6] One example to influence policy making is to employ lobbyists or support industry associations.

3.2.2 Programmatic measures

Programmatic measures enforce user behavior via software functions. They are more costly to develop and maintain because developers must design and integrate them into the software. These measures can restrict user inputs, which reduces the flexibility of user interfaces and may lead to bad user experiences. Some measures aim to influence user inputs by suggesting existing information (e.g., an autocomplete function that suggests product names). The main advantage of programmatic measures is that they are not or less dependent on a user's willingness or capability to comply with a policy, practice, or instruction. They provide reasonable complementary solutions for awareness measures. Table 9 outlines example factors for this type of measure. The second version of this deliverable will update the table.

Quality Characteristics	Quality Factors	Description
Accuracy	Syntactic defects	The syntactic problem is a problem of linguistic processing. It concerns how an author allocates roles such as subject and object in sentences and how they bind different meanings together.
Accuracy	Typographical errors	This factor means mistakes (such as a misspelled word) in a typed or printed text.
Understandability	Presence of acronyms	Unresolved acronyms make it difficult for readers to comprehend information.

Table 9. Example factors influenced by programmatic measures

3.2.3 Organizational measures

Programmatic measures can be too costly or restrictive for some complex use cases. In these cases, the producer can apply measures that rely on instructions, employee training, and creating organizational units or roles to manage data quality. The measures aim to provide, organize and validate data to increase or maintain the information quality. Organizational measures can introduce new information quality problems because the involvement of employees (human-in-the-loop) and work instructions create new error causes. Table 10 outlines example factors for this type of measure. The second version of this deliverable will update the table.

Quality Characteristics	Quality Factors	Description
Availability	Employee’s awareness of information existence	Employees may not be aware that needed information exists in their organization.
Accuracy	Training data sample size for machine learning	Small sample sizes may misrepresent the population. Trained models may be less accurate.
Accessibility	Access permission	Information retrieval requires permission.

Table 10. Example factors influenced by organizational measures

4. Conclusions

This document provides a guideline for managing the quality of production data. It establishes a conceptual basis by introducing several concepts, such as data and information, data life cycle, information needs, data and information quality, and production system levels. The guideline uses the Plan-Do-Study-Act (PDSA) cycle and focuses on the Plan and Do steps. Section 3.1 outlines an information flow analysis for producers to understand which data quality factors the organization must manage. Section 3.2 suggests three types of measures to manage data quality factors. Awareness measures aim to raise awareness of data quality issues and factors among employees. They require the least effort but are also not very reliable unless strictly controlled. Programmatic measures are functions in software that force users into behavior that ensures high data quality. Examples are input validations and auto-complete. These measures are much more reliable but may be costly to implement. Organizational measures cover complex cases where other measures are not feasible. They focus on larger-scale organizational activities (e.g., work instructions, training, and new roles) to promote behavior that minimizes data quality issues.

The proposed activity framework in **Section 3 fulfills the first goal** for task 3.1, i.e., providing key activities to manage data quality in manufacturing. This deliverable's second version will revise the activities based on insights from the use cases. The **second goal** concerns creating a methodological connection to other tasks (mainly in WP3). This document does **not yet meet** this goal, but the second version of this deliverable will contain an additional section to explain how other i4Q solutions and the activity framework align.

The **next steps** for Task 3.1 and the second version of this deliverables are:

- Perform the information flow analysis above with at least two end-users to update the list of quality factors. This analysis will focus on a few specific data quality issues to reduce the analysis complexity.
- Update the quality characteristics, factor groups, and quality factors in QualiExplore.
- For specific factors, we will identify control activities, suggest procedures to perform them, monitor / analyze the results, and take corrective actions from the measures of the activity framework and i4Q solutions (to meet the second goal above).

References

Braga-Neto U (2020) Fundamentals of Pattern Recognition and Machine Learning. Springer International Publishing, Cham

Case DO, Given LM (2016) Looking for information: A survey of research on information seeking, needs, and behavior, 4th edn. Emerald, Bingley

Chen J (2021) Corporate Governance. https://www.investopedia.com/terms/c/corporategovernance.asp

Choo CW (2000) Information management for the intelligent organization: The art of scanning the environment, 2nd edn. ASIS monograph series. Information Today, Medford, NJ

Frické M (2009) The knowledge pyramid: a critique of the DIKW hierarchy. Journal of Information Science 35:131–142. https://doi.org/10.1177/0165551508094050

ISO 25024:2015 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality. International Organization for Standardization

ISO 8000-1:2022 (2022) Data quality — Part 1: Overview. International Organization for Standardization

ISO 9000:2015 (2015) Quality management systems — Fundamentals and vocabulary. International Organization for Standardization

ISO 9001:2015 (2015) Quality management systems — Requirements. International Organization for Standardization

ISO/IEC 11179-1:2015 (2015) Information technology — Metadata registries (MDR) — Part 1: Framework. International Organization for Standardization

ISO/IEC 2382:2015 (2015) Information technology — Vocabulary. International Organization for Standardization

ISO/IEC 25012:2008 (2008) Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. International Organization for Standardization

Krcmar H (2015) Informationsmanagement. Springer Berlin Heidelberg

Lenzerini M (2002) Data integration. In: Abiteboul S, Kolaitis PG, Popa L (eds) Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02. ACM Press, New York, New York, USA, p 233

Liu L, Chi LB (2002) Evolutional Data Quality: A Theory-Specific View. In: MIT (ed) Seventh International Conference on Information Quality (IQ 2002), pp 292–304

N.a. (2022) PDSA Cycle. https://deming.org/explore/pdsa/. Accessed 20 June 2022

NIMBLE Consortium (2020) Collaboration Network for Industry, Manufacturing, Business and Logistics in Europe. https://www.nimble-project.org/. Accessed 20 June 2022

Nyhuis P (ed) (2008) Wandlungsfähige Produktionssysteme: Heute die Industrie von morgen gestalten. PZH, Produktionstechn. Zentrum, Garbsen

Taylor RS (1968) Question-Negotiation and Information Seeking in Libraries. CRL 29:178–194. https://doi.org/10.5860/crl_29_03_178

Westkämper E (2008) Fabriken sind komplexe langlebige Systeme. In: Nyhuis P (ed) Beiträge zu einer Theorie der Logistik. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 85–107