

RECOMMENDED PRACTICE

DNVGL-RP-0497

Edition January 2017

Data quality assessment framework



FOREWORD

DNV GL recommended practices contain sound engineering practice and guidance.

© DNV GL AS January 2017

Any comments may be sent by e-mail to rules@dnvgl.com

This service document has been prepared based on available knowledge, technology and/or information at the time of issuance of this document. The use of this document by others than DNV GL is at the user's sole risk. DNV GL does not accept any liability or responsibility for loss or damages resulting from any use of this document.

CHANGES – CURRENT

This is a new document.

CONTENTS

Changes – current.....	3
Section 1 Executive summary.....	6
1.1 Introduction.....	6
1.2 Data quality assessment.....	6
1.3 Organizational maturity assessment.....	7
1.4 Data quality risk assessment.....	9
Section 2 General.....	10
2.1 Introduction.....	10
2.2 Objectives.....	10
2.3 Scope.....	10
2.4 Related topics and framework foundation.....	11
2.5 Audience.....	12
2.6 Abbreviations.....	12
2.7 Terms and definitions.....	12
Section 3 Data quality assessment overview.....	14
3.1 Introduction.....	14
3.2 Define scope.....	14
3.3 Data exploration and profiling.....	15
3.4 Data quality assessment.....	15
3.5 Organizational maturity assessment.....	15
3.6 Data quality risk assessment.....	16
3.7 Risk based data quality improvement.....	16
Section 4 Data quality assessment.....	18
4.1 Introduction.....	18
4.2 ISO 8000-8.....	18
4.3 Profiling.....	20
4.4 Evaluating data quality metrics.....	21
Section 5 Data quality maturity areas.....	25
5.1 Introduction.....	25
5.2 Governance.....	25
5.3 Organization and people.....	25
5.4 Processes.....	25
5.5 Requirement definition.....	25

5.6 Metrics and dimensions.....	25
5.7 Process efficiency.....	26
5.8 Architecture, tools, and technologies.....	26
5.9 Data standards.....	26
Section 6 Organizational maturity levels.....	27
6.1 Introduction.....	27
6.2 Level 1 - initial.....	27
6.3 Level 2 - repeatable.....	27
6.4 Level 3 - defined.....	27
6.5 Level 4 - managed.....	28
6.6 Level 5 - optimized.....	28
Section 7 Data quality process maturity assessment.....	29
7.1 Introduction.....	29
7.2 Maturity score.....	29
7.3 Evaluating framework elements.....	30
Section 8 Data quality risk assessment.....	35
8.1 General.....	35
Section 9 Security.....	39
9.1 General.....	39
Section 10 References.....	41
10.1 List of references.....	41
Changes – historic.....	42

SECTION 1 EXECUTIVE SUMMARY

1.1 Introduction

In the modern digitized world, most advanced industrial operations are dependent on information systems for control and analysis. Data is increasingly being considered a valuable asset, of equal worth to physical assets, and considerable costs are involved in collecting, storing, and acting upon the data. As with physical assets, the quality of the data is a prerequisite for ensuring reliable operations. Additionally, high data quality must be ensured to enable reuse of data and to enable analytics on historical data. Information and analytics-driven organizations, with no traditional physical operational commitment, rely solely on high quality data to stay competitive. Digitalization in traditional industries blurs the boundary between running traditional versus digital business operations. Actors are moving towards operating partly or fully as information and analytics-driven.

Data value chains are common in industry, production and business operations. Data is born, follows a value chain, and is then refined and prepared for several different tasks. Thus, the user or system utilizing the data does not necessarily have knowledge of the data origin, quality level, weaknesses, legal or contractual obligations, semantics, changes in the system capturing data, the context in which the data was born etc.

In order to ensure both reliable operations and valid analytics, it is important that the data quality is assessed and continuously monitored for all critical systems and services. Organizations should define data quality policies, and processes should be in place to support these policies. The requirements for data quality and dataset definitions must be clearly stated, and measurement points should be implemented in order to verify compliance with requirements. Ideally, the measures should be in effect across the entire organization to ensure optimization and to avoid data quality assessment being performed in silos.

The framework outlined in this recommended practice (RP) consists of three main parts, each described in separate sub-sections below.

1.2 Data quality assessment

Firstly, the RP describes a generic tool for evaluating data quality and provides concrete suggestions on how to obtain data quality measurements according to metrics and defined criteria.

ISO 8000-8 *Information and data quality: Concepts and measuring* defines three categories for data quality measurements; syntactic, semantic, and pragmatic, and provides a foundation for measuring information and data quality. Information and data quality are defined and measured according to the following categories:

- Syntactic quality: the degree to which data conforms to their specified syntax, i.e. requirements stated by the metadata.
- Semantic quality: the degree to which data corresponds with that which it represents, i.e. when a sensor measures 72 °C, the actual temperature should also be 72 °C at the point of measurement.
- Pragmatic quality: the degree to which data is appropriate and useful for a particular purpose.

In the Internet of things (IoT) and sensor systems, data feeds for operations may provide data of higher quality than data feeds to analytics, as the former typically requires fewer steps from data generation to data use. Data collected for analytics are frequently processed several times by logging systems and data transformations, processes that could adversely affect the data quality.

ISO 8000-8 suggests that data will either fulfil requirements - or not. To be able to classify data as bad or good, as illustrated in [Figure 1-1](#), strict data quality requirements must be in place.

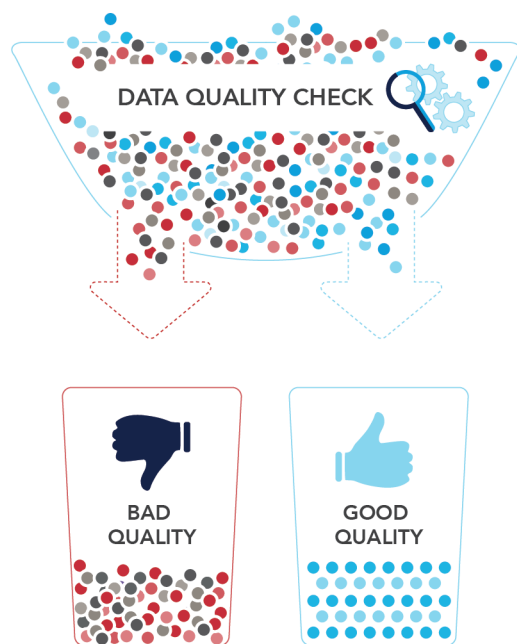


Figure 1-1 Data quality check outcome

Figure 1-1 shows the ISO 8000-8 principle of evaluating data as correct (good) or incorrect (bad).

1.3 Organizational maturity assessment

The second part of the RP is a framework that aims at encompassing all processes, capabilities, and governance that are required for ensuring high data quality. Organizations organize, build, operate, and monitor their data quality activities and capabilities according to their internal needs, legacy, and goals. Regardless of the detailed implementation of data quality activities, organizations have different levels of maturity and focus on data quality. Process quality and capability quality are measured according to 5 maturity levels, ranging from ad hoc and reactive, to controlled and proactive. The initial maturity level (level 1) indicates a low data quality skillset and lack of governance and data requirement definitions. The optimal maturity level (level 5) is typical for organizations with a high degree of data quality awareness, continuous improvements, and well-governed, enterprise-wide data quality processes. Intermediate levels represent a gradual transition between the two endpoints.

The data quality maturity framework consists of elements, maturity levels, and evaluation criteria. The elements describe governance, processes, technologies, capabilities, and activities that must be implemented to support data quality. Detailed evaluation criteria are given for each framework element at each maturity level. The maturity levels provide a measuring device for the elements, and the data quality assessments provide a means to verify and validate each dataset. There is no 1:1 correlation between organizational maturity and data quality, and organizations could, theoretically, have a high score for maturity and a low one for data quality, or vice versa. However, high maturity levels are commonly associated with high data quality.

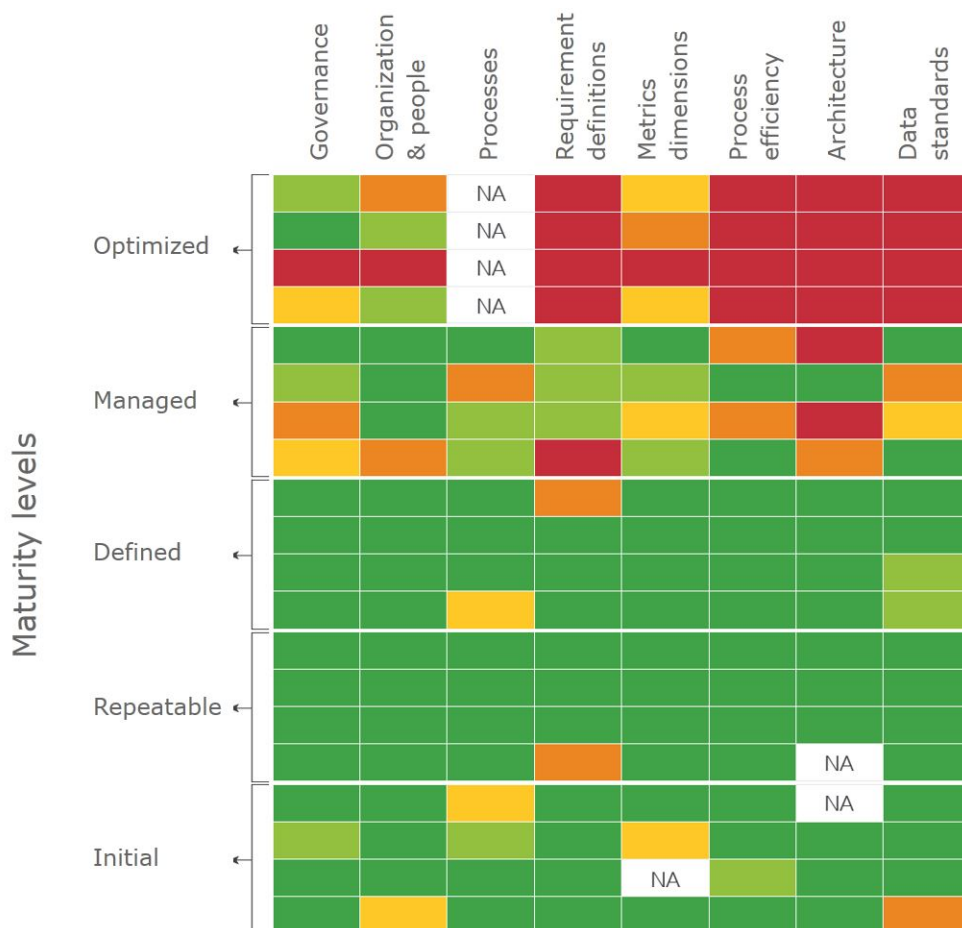


Figure 1-2 Sample outcome from a maturity assessment

The maturity level should be determined for organizations collecting, using, and sharing data in order to assess the reliability of the data and the ability to respond to data quality incidents in a predictable and repeatable manner. Limited or lack of data sharing or data integration is typical for organizations below level 3 (defined). To comply with level 3, organizations must define data quality requirements, and data standards must be established, in addition to governance and communication channels. At level 3, data quality can be measured according to well-established standards, such as ISO 8000-8. At level 4, data quality can improve continuously by means of feedback processes and change management. The data quality maturity level is determined by data analysis, document reviews, and interviews with key personnel. The results can be displayed as a matrix showing the maturity level for each framework element. If the target level is known, then a gap analysis can be performed to determine the roadmap to reach the desired level.

Based on an organization's maturity level, different data quality measurements are chosen to evaluate the data itself. At levels 1 and 2, where data quality definitions and data quality processes are absent or incomplete, profiling-based and reengineering techniques are commonly used to assess data quality in an informal manner. As the maturity increases to level 3 and above, more stringent data quality standards, such as ISO 8000-8, can be deployed. ISO 8000-8 necessitates complete definitions for both metadata and the conceptual model.



1.4 Data quality risk assessment

The impact of data quality on operations and/or analytics should be assessed using common tools for risk analysis, such as bowtie models, risk matrices, and fault tree analysis. Relevant contexts should be established and any risk causes, barriers, events, and consequences should be defined that are relevant to the context. Barriers upstream of the initiating data quality related events are proactive and usually implemented at the higher maturity levels, whereas downstream barriers are implemented as reactive measures for damage control. Outputs from the risk analysis will provide guidance for prioritizing activities to improve data quality.

Information security is tightly coupled to data quality and each of the main elements of information security - availability, confidentiality, and integrity affect elements of data quality. A high level of maturity of data quality is generally associated with higher levels of information security.

SECTION 2 GENERAL

2.1 Introduction

Data quality assessment is a method to verify that data meets the implicit or explicit expectations of users or systems utilizing the data. The data quality measurement score shows how the data meets these expectations. The assessment of data quality-related processes and capabilities is measured as organizational maturity according to set criteria. These two different measurements are input to data quality risk and improvement activities. Risk assessment uses these measurements as input for obtaining a risk picture of data usage. Measured and documented data quality and maturity levels, within risk tolerance thresholds, are viewed as prerequisites to safe, sustainable, and efficient business operations.

2.2 Objectives

The objective of this recommended practice (RP) is to provide the industry with a systematic approach and framework for the assessment of data quality. Assessment of both (i) data quality per se and (ii) data quality-related processes and capabilities are considered in this RP. The assessment methodology is valuable input for organizations on how to establish best practices for achieving satisfactory data quality.

2.3 Scope

Data and information should both be considered as assets in themselves, and as parts of, and prerequisites for, the operation of physical assets like ships, oilrigs, power grids etc. Based on this view, our framework is built on standards and best practices in asset management in general, and data management in particular. In relation to standards dealing with asset management, such as ISO 55000, this RP views data in the scope of a business context of a portfolio of assets to be managed. The portfolio of datasets or data repositories is governed by an asset management system, and specifically, the data quality management is part of that system. The asset management system for data is part of the organization asset management (often called governance level). Purpose of the latter is coordinating activities and defining goals and risk tolerances, in order to realize value from information as an asset.

Corporate governance, as specified by policies, guidelines, and management systems, governs the overall criteria for performing business. Information governance guides information management at the enterprise level and supports all operational, legal, environmental, and regulatory requirements. Data management ensures that important data assets are formally managed throughout the enterprise and that information governance goals are achieved. This RP covers assessment of data quality, which is an important subset of data management.

Data quality governance policies should be defined as an integral part of corporate governance, guided by management systems. Risk assessments should be performed at several levels of governance in order to define critical issues for both data quality and information security. Some critical issues could have impacts on both quality and security; e.g., unauthorized access to data could result in compromised data values and, in contrast, ill-defined data quality processes could jeopardize security.

This document does not contain details on how to check data quality in each context, source code examples for developers, etc. This recommended practice:

- defines a framework for how to measure whether data quality in data sources is in accordance with criteria relevant for the given context
- defines a framework for how to measure the maturity of an organization that is responsible for ensuring adequate data quality for a given purpose, and
- a risk analysis approach is used to analyse risk, and prioritize mitigation activities according to the risk score and risk tolerance.

2.4 Related topics and framework foundation

Data quality is part of the broader topic of data management, which, in turn is part of information governance. Information governance is the set of multi-disciplinary structures, policies, procedures, processes, and controls that are implemented in order to manage information at an enterprise level. Information governance is supporting an organization's immediate and future goals, and ensuring that various requirements (regulatory, legal, risk, environmental and operational), are met. Information governance should determine the balance between two potentially divergent organizational goals: extracting value from information and minimizing the potential risk from using information.

Data management ensures that important data assets are formally managed throughout the enterprise and that information governance goals are achieved. Data management establishes and utilizes processes, controls, and technologies that operate on data and enable the data maintenance and data value chain in the enterprise to be compliant with policies and governance directions.

In addition to focus on data quality, data management also covers management of:

- data storage, retention, life cycle, legal compliance and intellectual property rights, value chains, metadata, business vocabularies, information risks, and information security etc.

Information and communications technology (ICT) management covers the life cycle and operative procedures and policies for the IT-systems used to capture, use, store, and monitor data and its usage.

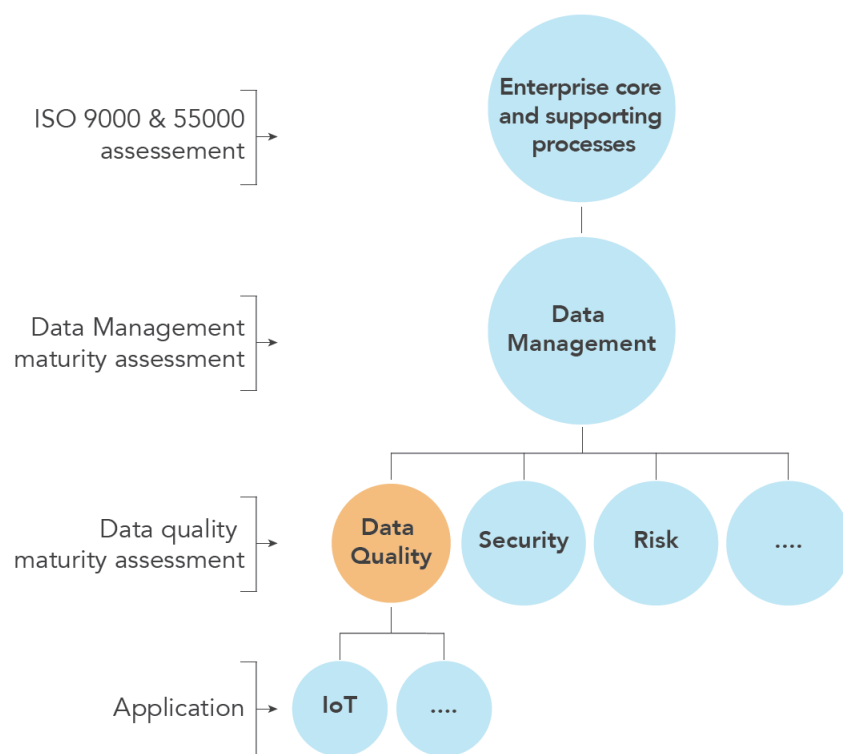


Figure 2-1 Data management document hierarchy

Figure 2-1 illustrates a set of RP documents needed for assessment of sub-topics in data management. The same hierarchy can be used for DNV GL improvements services. The full title of the document illustrated in the IoT circle in Figure 2-1 is *Data Quality Assessment - For Sensor Systems and Time Series Data /16/*.

ISO 8000-8 *Information and data quality: Concepts and measuring* /1/, is used as a core reference in this RP for detailed verification and validation of data quality. The two data quality maturity models by Loshin /7/ and CMMI /6/, and W3C *data on the web, best practice* /5/ are used as the basis for maturity levels and framework elements. Relevant risk assessment methodology is described in ISO 31000 *Risk management* /4/. Information security is covered in DNVGL-RP-0496 *Cyber security resilience management for ships and mobile offshore units in operation* /14/ and ISO 27000 *Information security management* /3/.

2.5 Audience

The target audience for this document includes DNV GL customers, DNV GL consultants performing assessments, and the data quality community.

The RP covers data quality assessment services for external or internal data providers, as well as verification of digital assets such as sensor data used for digital classification.

2.6 Abbreviations

Table 2-1 Abbreviations

<i>Abbreviation</i>	<i>Description</i>
IoT	Internet of Things
IPR	intellectual property rights
RP	recommended practice
SLA	service level agreement
W3C	World Wide Web Consortium

2.7 Terms and definitions

Table 2-2 Terms and definitions

<i>Term</i>	<i>Definition</i>
conceptual model	data model that represents an abstract view of the real world SOURCE: ISO/IEC 11179-1:2004
data	reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing SOURCE: ISO 8000-8, 2015, referencing source ISO/IEC 2382:2015
data element	unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes
data identifier	unique identifier for an administered item within a registration authority SOURCE: ISO/IEC 11179-1:2004
data management	data management ensures that important data assets are formally managed throughout the enterprise and that information governance goals are achieved It establishes and utilizes processes, controls, and technologies that operate on data and enable compliance with policies and governance directions of data maintenance and the data value chain in the enterprise.

<i>Term</i>	<i>Definition</i>
data quality	measurement to which degree data meet the implicit or explicit expectations and requirements of users or systems utilizing the data Information and data quality are defined and measured according to syntactic, semantic and pragmatic quality. Syntactic and semantic quality is measured through a verification process, whereas pragmatic quality is measured through a validation process.
data repository	a population of records of one of more kinds
dataset	a population of records of the same kind
data value	content of a data element
information	knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, which have a particular meaning within a certain context SOURCE: ISO 8000-8, 2015, referencing source ISO/IEC 2382:2015. SOURCE: ISO 8000-8, 2015, referencing source ISO/IEC 2382:2015
information governance	the set of multi-disciplinary structures, policies, procedures, processes, and controls that are implemented to manage information at an enterprise level, supporting an organization's immediate and future goals and ensuring achievement of different requirements, which may be regulatory, legal, risk, environmental, or operational
metadata	data that defines and describes other data SOURCE: ISO/IEC 11179-1:2004
pragmatic quality	the degree to which data is appropriate and useful for a particular purpose SOURCE: ISO 8000-8, 2015
record	a collection of data element values with an ID
reference data	data used to classify or categorize other data. (e.g., nation codes, metric system for weights) SOURCE: The Data Management Body of Knowledge
semantic quality	the degree to which data corresponds with that which it represents SOURCE: The Data Management Body of Knowledge
syntactic quality	the degree to which data conforms to their specified syntax, i.e., requirements stated by the metadata SOURCE: ISO 8000-8, 2015
validation	confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled SOURCE: ISO 9000:2015
verification	confirmation, through the provision of objective evidence, that specified requirements have been fulfilled SOURCE: ISO 9000:2015

SECTION 3 DATA QUALITY ASSESSMENT OVERVIEW

3.1 Introduction

This section gives an overview of the data quality assessment and improvement process. The process is shown in [Figure 3-1](#). The following sub-sections provide more in-depth details of the steps in the figure below.

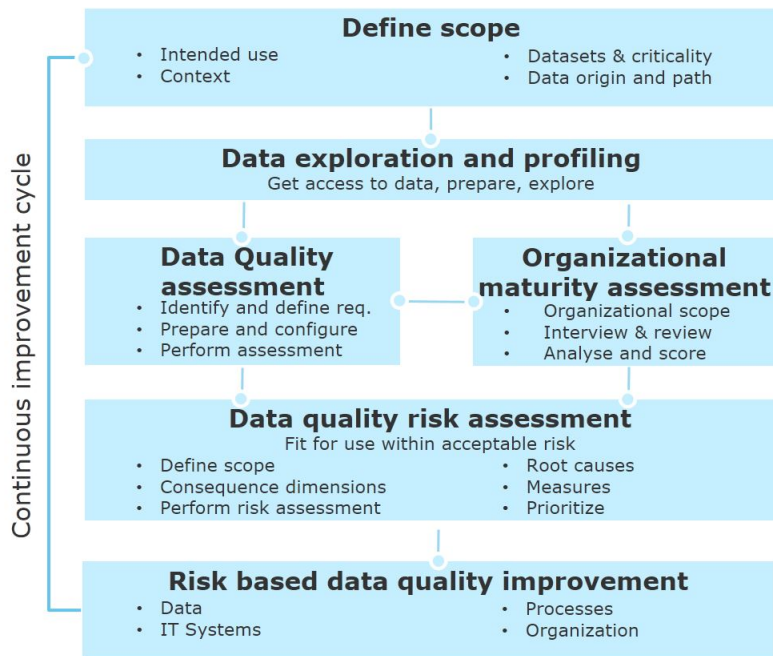


Figure 3-1 Data quality process

3.2 Define scope

Defining the scope for the data quality assessment process includes limiting the universe of discourse and making a clear, concise description of the intended uses of the data and its sources. Part of this process is to give a high-level description of the relevant datasets, their criticality related to intended use, the data origin, and the path and steps that the data has been following.

Data quality assessment can be performed on:

- streaming data, batches, or data at rest
- structured or non-structured data
- large or small volumes of data
- siloed or distributed data
- data value chains
- digital twins, simulation models, semantic models, ontologies, or taxonomies.

The method described here should be tailored to the context, types of data, and scope.

3.3 Data exploration and profiling

In order to be able to perform data quality assessments, the rules for data and assessment must be connected to the same logical and computational site. Based on the scope definition, a list of datasets can be identified. To be able to continue, access to these data is necessary and the data must be prepared for exploration and profiling.

The first iteration in the data quality assessment process typically has no or minimal requirements defined, and data exploration and profiling assists in refining requirements. Organizations at maturity level 3 and above should have a good set of requirements established. All users and domain experts should be involved in establishing the requirement definitions. Data quality requirements should focus on those criteria that are relevant for the intended use and be based on schema, metadata, domain models, business rules, any relevant policies or governance documents, algorithm requirements, data origins, and life cycles, etc.

This topic is further described in [Sec.4 \[3\] Profiling](#).

3.4 Data quality assessment

Data quality assessments are typically performed both bottom-up and top-down. The bottom-up approach utilizes profiling tools and schema inspections to perform a generic and usage-agnostic assessment. The bottom-up approach will reveal indicators of potential areas of data inconsistency. However, due to the generic nature of the method, this approach is also prone to detect false positives. The top-down approach will involve domain experts and actual usage scenarios to detect inconsistencies. Although this method does not typically result in false positives, it will not lend itself easily to automation and hence might not be as conclusive as the bottom-up approach. Normally, the bottom-up assessment provides valuable input to the top-down assessment, and hence both are required to perform an exhaustive evaluation.

Performing an initial iteration is recommended in order to validate input data flow, map data paths and all transformations, identify enhancements and refinements on data, collect and use metadata and schemas involved, and document this as part of the data quality assessment. The next iteration will detect any relationships between data feeds. These could include: (i) a sensor measuring power supply to a pump can be correlated to sensor data measuring performance of a pump, or (ii) starting an engine, resulting in a rise in engine oil temperature being detected by sensors. Subsequent iterations can assess whether the data quality is sufficient for the algorithms using them.

Algorithms are known to have variable vulnerabilities to data quality issues. Some methods to measure the impact of data quality issues on data-mining algorithms have been described in the literature /12/ and /13/. If simulations and/or digital twins are used, these models should also be quality assured. For this purpose a digital twin can be viewed as a dataset in the context of the data quality method. The first assessment provides a baseline for measurements and the improvement cycle.

ISO 8000-8 Data quality, defines three categories for data quality measurements; syntactic, semantic, and pragmatic. Information and data quality are defined and measured according to these categories.

For organizations with well-defined requirements, the assessment will tend towards that of the assessment model for ISO 8000-8 shown in [Figure 4-1](#). In this case, the data quality dimensions are categorized according to ISO 8000-8 and the appropriate methods are employed:

- automatic syntax and integrity checks for syntactic quality
- correlation with reference models and sampling techniques for semantic quality
- algorithm sensibility for data quality issues, user feedback, and focus groups for pragmatic quality.

This topic is further described in [Sec.4 Data quality assessment](#).

3.5 Organizational maturity assessment

Maturity level should be determined for organizations collecting, using, and sharing data such that the reliability of the data and the ability of the organization to respond to data quality issues in a predictable and repeatable manner can be assessed. In order to comply with level 3, organizations must define data quality requirements, and data standards, as well as governance and communication channels, must be established.

At this level, data quality can be measured according to well-established standards, such as ISO 8000-8, and the data quality can improve continuously by means of feedback processes and change management.

The data quality maturity level is determined by data analysis, document reviews, and interviews with key personnel. If the target maturity level is known, then a gap analysis can be performed to determine a roadmap to reaching the desired level.

This topic is further described in [Sec.5 Data quality maturity areas](#), and [Sec.6 Organizational maturity levels](#).

3.6 Data quality risk assessment

Pragmatic assessment evaluates whether data is fit for use and requests user feedback. In addition, a more holistic and systematic approach is needed. The results from the data quality assessment must be validated and evaluated by both domain experts and data users. The relevance of the findings and the potential impact on business should both be carefully analysed. Risk analysis tools can be employed to evaluate and prioritize mitigation activities according to the risk score and risk tolerance. If improvements are needed, the maturity of the organization should be evaluated to determine where and how improvements could be targeted.

Fitness for use and risk score should be considered for the relevant usage and business context. Output from the risk analysis will be used as input to improvements. For high risk scores, an organizational maturity assessment is also recommended. This will give indications of possible root causes, plus a baseline of the organizational capability and maturity related to data quality processes and performance.

This topic is further described in [Sec.7 Data quality risk assessment](#).

3.7 Risk based data quality improvement

The results of the data-quality risk assessment are measures, activities, projects etc. that could be actioned as approaches towards improvement of the current situation.

The goals of the risk-based data quality improvement process are: to improve performance in digital processes, to gain business benefits, to harvest opportunities, and to reduce costs. In some cases, the quality can be improved by enhancement and cleansing, but at-source improvements provide a more robust process. In order to facilitate this improvement process, several key players in the organization must cooperate. These include: system owners, data collectors, data owners, process owners, and policy/governance owners. The improvement process will be iterative and should yield improvements both in the process itself (maturity), but also in the data quality (DQI). The baseline provided by the previous assessment, can be used to measure both the improvement in data quality and the efficacy of the data quality processes.

The data quality service measures the quality level and evaluates values or records as being within or outside the defined criteria. Trust in the data quality service is directly influenced by its ability to perform correct evaluations. For example, in virus detection terms, a false positive means that an object is incorrectly identified as a virus, whereas the term false negative is used when a virus goes undetected. The same challenge is valid for data quality evaluation. The business impacts of false negatives and false positives may be substantial, and considerable efforts should be invested to reduce such occurrences. [Figure 3-2](#) illustrates this problem. False positives are shown as red circles in the good quality zone, whereas false negatives are green circles in the uncertain zone.

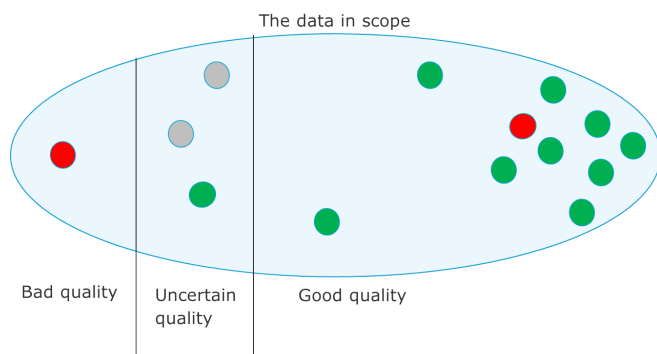


Figure 3-2 False positive and false negatives

The combination of data quality and organizational maturity assessments will support organizations with the following:

- an increase in the good data quality area as shown in [Figure 3-2](#).
- a reduction in the uncertain area and the bad quality area
- a decrease in the number of false positive and false negative measurements.

SECTION 4 DATA QUALITY ASSESSMENT

4.1 Introduction

The metrics and dimensions used for data quality assessments should provide a coherent and complete evaluation framework for data quality. Several such frameworks are provided in the literature, however, this RP follows the definitions provided by ISO 8000-8 /1/.

4.2 ISO 8000-8

ISO 8000-8 defines three categories for data quality measurements: syntactic, semantic and pragmatic. It provides a foundation for measuring information and data quality.

- Syntactic quality: the degree to which data conforms to its specified syntax, i.e., requirements stated by the metadata. Metadata in this context is, e.g., legal values, data types, referential integrity such as links between data parts, business vocabulary, and any defined business rules.
- Semantic quality: the degree to which data correspond to what it represents, i.e. when a sensor measures 72 °C, the actual temperature should also be 72 °C at the point of measurement, and if not, there is some amount of semantic error. Similarly, if a passenger list shows 162 passengers, there should also be exactly 162 passengers in the actual real world situation.
- Pragmatic quality: the degree to which data is suitable and useful for a particular purpose. It validates data users' perceptions of fitness for purpose.

We recommend that fitness for use is evaluated in two steps. First, an initial evaluation corresponding to the ISO 8000-8 regarding whether data is fit for use, and, afterwards, a risk analysis investigating selected consequences and generating risk matrixes for data quality-related risks.

Syntactic and semantic measurements are data quality verifications, whereas pragmatic measurements are data quality validations:

- Verification is the evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition.
- Validation is the assurance that a product, service, or system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability with external customers. This often involves acceptance and suitability tests with external parties.

As indicated in [Figure 4-1](#), syntactic quality is measured as conformance with the schema, metadata, and any defined business requirement definitions, rules, and vocabulary. Syntactic verification is normally fully automated and the entire dataset is considered. Semantic data quality measurements verify conformance between the data and the entity or object that the data represents. Verification uses a conceptual model, which is a digital model or representation of the actual entity. Semantic verification is normally performed by statistical sampling methods in order to achieve the desired percentile. In some cases, trusted surrogates may be used. Trusted surrogates are reference data that is sufficiently trusted to be a substitute for real world measurements. Dun & Bradstreet databases on more than 265 million companies across 200 countries is an example of a trusted surrogate, for which company-related data could be verified according to the register rather than by direct enquiries. Pragmatic data quality validation represents users' perceptions of the data and is normally measured by user groups, questionnaires, and feedback loops.

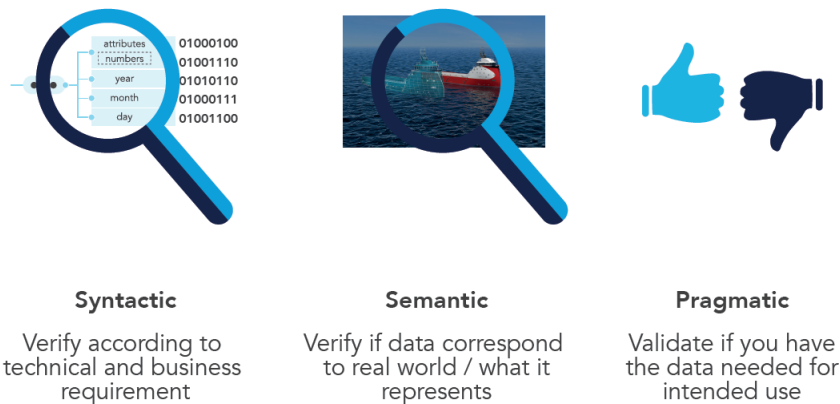


Figure 4-1 ISO 8000-8 assessment model

ISO 8000-8 requires both a complete schema/metadata definition and a conceptual model in order for syntactic and semantic verifications to be performed. Schema and conceptual models are important input for establishing data quality requirements and measurement criteria. It also requires a complete data quality requirement definition as a basis for all measurements. These requirements imply that a minimum maturity level of 3 is achieved, prior to performance of a full ISO 8000-8 compliance verification. Pragmatic validation and parts of syntactic validation could be performed at lower maturity levels and provide valuable input to the implementation of data quality processes.

Figure 4-2 illustrates one view of the business model for DNV GL data quality service based on this RP.



Figure 4-2 DNV GL data quality service

4.3 Profiling

4.3.1 Introduction

Data profiling is a valuable and commonly used method to investigate potential quality issues empirically in datasets with little or no metadata available. The method is an efficient means to reengineer metadata for a dataset and to detect possible inconsistencies, such as illegal values, outliers, or other statistical anomalies. Data profiling is often used in the early stages of data activities, such as migrations, integrations, and analytics, in order to be able to anticipate whether actions might be necessary to address the quality issues.

Data profiling is commonly used at the early levels of data quality maturity, where little or no metadata definitions exist (level 1). As data quality requirements are developed and the project progresses to higher levels of maturity, the data quality assessments will shift to more formal definitions, such as those provided by ISO 8000-8. Data profiling can still be useful at the higher maturity levels, however, it is commonly regarded as an ad hoc and reactive process, possibly used as input to data cleansing. The higher data quality maturity levels will pursue proactive processes where data requirements will be communicated and governed to minimize both profiling and cleansing.

4.3.2 Data profiling techniques

Several statistical and counting techniques are used to profile datasets, both for single datasets and for connected datasets. Single datasets commonly employ the assessments below.

Table 4-1 Profiling techniques

Profiling techniques		
Generic data types	Foreign key flag	High value count
Declared data types	Null flag	Mean value
Declared maximum length	Low value	Median value
Declared precision	High value	Minimum, maximum, mean, median string length
Primary key flag	Low value count	Distinct count, percentage
Null count, percentage	Zero count, percentage	Blank count, percentage
Pattern count	Valid count, percentage	Distinct valid count, percentage

The profiling techniques listed in [Table 4-1](#) is part of existing well known profiling tools /18/.

If key definitions are available, then cross-table validations can also be performed. The criteria in [Table 4-2](#) can be used for evaluation of referential integrity across tables/ dataset A and B.

Table 4-2 Cross-table profiling

Cross-table profiling		
Number of records in A	Percentage of records in A also found in B	Number of records in B found in A
Number of records in A not in B	Number of records in B	Percentage of records in B found in A
Percentage of records in A not in B	Number of records in B, but not in A	Number of records matching

<i>Cross-table profiling</i>		
Number of records in A also found in B	Percentage of records in B but not in A	

4.3.3 Data profiling and ISO 8000-8

Data profiling techniques generally provide candidates for syntactic and semantic verifications according to ISO 8000-8. The main differentiator between ISO 8000-8 and data profiling is the strict requirement for data definitions in ISO 8000-8. Data profiling can be performed prior to agreement on data definitions and data quality requirements and hence data profiling is particularly useful at low maturity levels where metadata could be missing or incomplete.

4.4 Evaluating data quality metrics

4.4.1 Introduction

Data quality metrics are commonly grouped into data quality dimensions. The data quality dimensions can be formalized and categorized according to ISO 8000-8 or they can be implemented in an ad hoc manner as a part of data profiling. Several visual devices exist to support the evaluation of the data quality metrics. At a minimum, the evaluation device should display the data quality index (DQI), core data quality metrics, and the data quality development indicator when historical data is available.

4.4.2 Dimensions of data quality

Within each of the data quality dimensions listed in [Table 4-3](#) we seek to define data quality requirements. The list is discussed and explained in the DNV GL guide "Data quality for sensor data and time series data" /16/.

Table 4-3 Data quality dimensions

<i>Data quality dimensions</i>	
<ul style="list-style-type: none"> — Accessibility — Accuracy — Actuality — Auditability — Authorization — Completeness — Conformance to domain models and conceptual model/ Semantic consistency — Conformance to metadata /schema — Consistency — Duplicates 	<ul style="list-style-type: none"> — Format/structural consistency — Integrity — Interoperability — Identifiers — Precision — Processability — Resolution — Timeliness — Traceability — Uncertainty

4.4.3 Data quality index

The DQI provides a normalized number that express the data quality as a decimal number, ranging from 0 to 1, where 0 is lowest quality and 1 represent the highest quality. If the total number of records evaluated is R_t and R_p is the number of records that passed the test, the DQI is defined as

$$DQI = R_p/R_t$$

resulting in a normalized decimal number ranging from 0 to 1. The DQI is normally displayed for each metric and also as an aggregated score for all metrics for the relevant data source.

4.4.4 Data quality dashboards

Data quality measurements can be displayed and communicated in several ways. Figure 4-3 below shows a dashboard for one solution with metrics: uniqueness, above minimum, below maximum, completeness, goodness and usable. A dashboard can also support drill downs on metrics to provide further insights into relevant data quality issues.

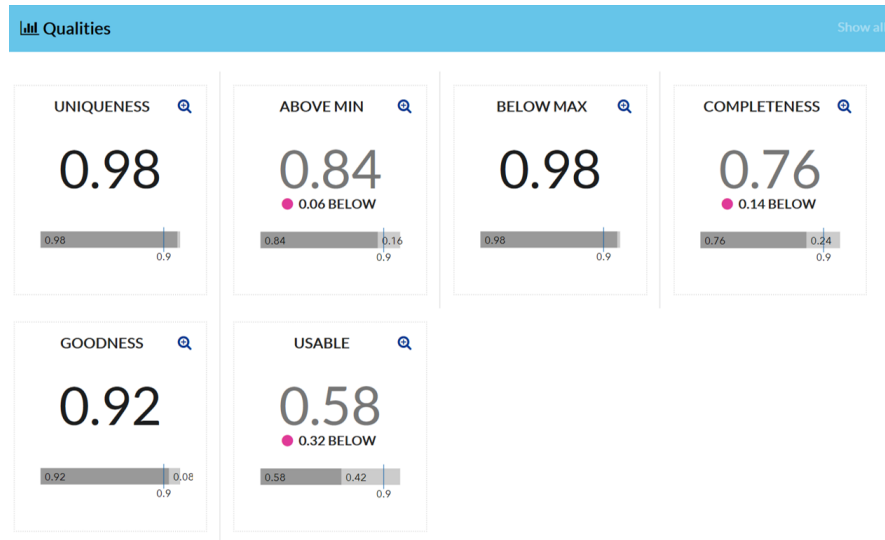


Figure 4-3 Data quality dashboard

Data quality metrics typically quantify the relationship between the total number of records in scope and the portion of records passing a criterion. When a defined threshold is not met, this can be indicated on the dashboard, e.g., by red flag.

In real life we often work with hundreds or thousands of sensors, and by grouping them according to a domain model, functionality or in line with digital twin models, a data quality score can be derived from several perspectives and drill down capabilities should exist.

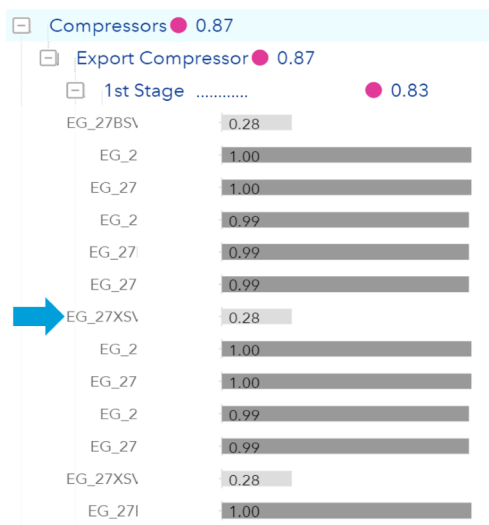


Figure 4-4 Data quality sensor list

4.4.5 Data quality index over time

Historical data quality assessments indicate improvements or deteriorations in data quality due to data quality processes and governance. The success of any data quality activities implemented can be evaluated from a historical perspective. Figure 4-5 shows development in completeness for a sensor with the value 0.28, marked with an arrow in Figure 4-4.



Figure 4-5 Data quality dashboard, index over time

Figure 4-5 shows an example of a sensor drill down data quality dashboard.

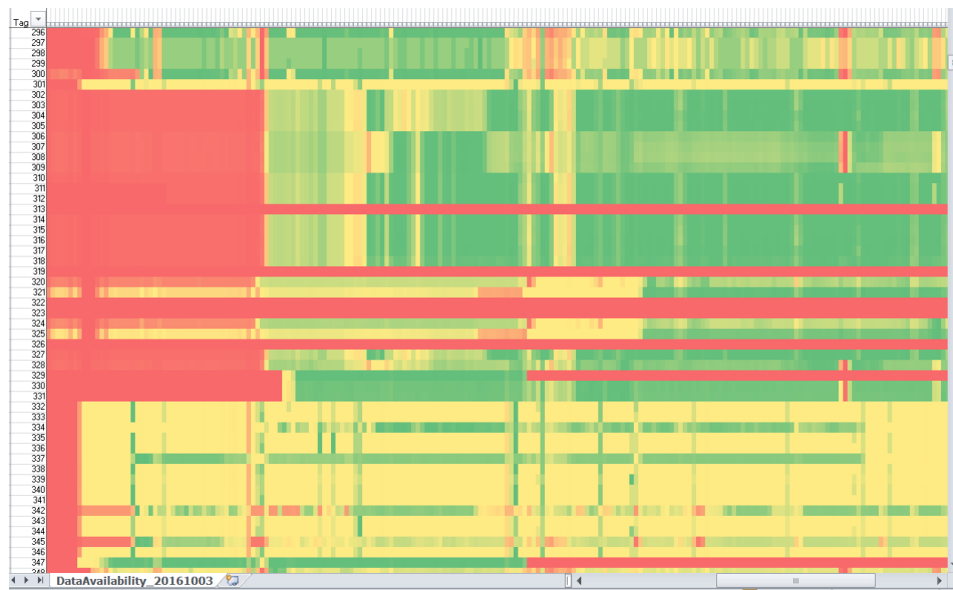


Figure 4-6 Data quality sensor heatmap

The heatmap above is part of a data quality study that was performed on several thousand sensors. Each vertical row represents a sensor. Each column represents a short time-window. The colour in each cell indicates the result of a data quality measurement.

SECTION 5 DATA QUALITY MATURITY AREAS

5.1 Introduction

The elements of the data quality framework described in this section encompass all the processes, technologies, and activities that are required to measure and improve data quality. Each element can be evaluated according to maturity level, and the combination of the particular element and level will indicate a roadmap to improved data quality.

5.2 Governance

Good data governance results in good data quality, and the subset of data governance that directly relates to data quality is part of the framework. Data governance defines the policies, processes, roles, and responsibilities that are required for continuous monitoring and improvement in data quality. Good guidelines must be established to ensure that any data quality issues are reported and addressed at the appropriate level. It is important that governance activities are supported by top-level management.

Data management ensures that important data assets are formally managed throughout the enterprise and that information governance goals are achieved. Data quality management is regarded as being part of the broader term of data management. Data quality is part of data management and data management is part of data governance.

5.3 Organization and people

Data management and data quality processes are performed by people in roles with defined responsibilities. The organization must empower employees to perform roles and fulfil responsibilities to achieve governance goals related to data quality. Employees must have the appropriate skill level, backing, mandate, and time for both reporting and remedying any discrepancies in the data effectively. Top leadership should be provided with the key performance indicators on data quality improvements and should be encouraged to publish the organization's data quality status openly.

5.4 Processes


Data quality processes are implemented to support the data quality policies, governance, and any service level agreements (SLA). The processes define workflows for both measuring and improving the quality and efficiency of the data management, as well as the quality of the actual data. The processes guide the participants on operational aspects of data quality management and on the actions that should be taken should quality issues be discovered. Processes for data quality operations and improvement have many parallels in software operations and improvements.

5.5 Requirement definition

Data requirement definition is the specification of business needs, value chain needs, and the technical properties that are to be met by data and datasets. The definitions for data requirements must be agreed upon by all stakeholders and for all relevant areas of data usage. The data requirements must cover all aspects of data integration, reporting, analytics, and operations. The requirements could be defined as an intrinsic part of the data model or could be rooted in other internal or external technical or business expectations. The requirements should be collated, documented, and matched with appropriate metrics and measurement devices.

5.6 Metrics and dimensions

The data quality dimensions provide a classification of data requirements. The dimensions could be adopted from ISO 8000-8 and classified according to syntactical, semantic, or pragmatic categories. A plethora



of other definitions exist, and several good benchmarks of the different definitions can be found in the literature. In this RP we provide some descriptions of the dimensions based on ISO 8000-8 and suggest adhering to the principles from ISO 8000-8. The data quality metrics define the measurements to be performed for the selected dimensions.

For maturity levels 1 and 2, the ISO 8000-8 metrics and dimensions are not appropriate. For these levels, data quality assessment should be based on techniques like data profiling, data reengineering, statistics, pattern analytics, and machine learning.

5.7 Process efficiency

Measuring the efficiency of the data quality governance, management, processes, and tools that have been implemented will provide both baselines and a device to assess the impact and improvements of the data quality activities. Performance data is collected, and trend lines used to measure the rate of continuous improvement. Activities are adjusted according to negative or positive trends. Audit trails and change logs are used to document all relevant data quality issues. Data quality dashboards communicate the status for data quality processes and data quality metrics to both users and management. The measurements provide input to root cause analysis and adjustments to current data quality practices.

5.8 Architecture, tools, and technologies

Relevant tools, technologies, and architecture are required to support the data quality framework. Data quality dashboards, verification algorithms, monitoring services, profiling software, auditing measures, and change logs all need coherent tools. In addition, modelling tools, reengineering tools, and integration tools are required for successful connection to raw data, refining data, calculating new data, and for verifying and/or reengineering any metadata issues. Tools, architecture, and infrastructure for data exploration, profiling, and quality analytics need to be in place and used.

5.9 Data standards

The term "data standards" describes, in this context, the collection of specifications and protocols used to capture, ingest, structure, describe, share, trace, and secure data during its life cycle and value chain. The main goals for these standards are supporting interoperability, usage, maintenance, and value creation. Standards for metadata, reference data, data representations, integrations, semantics, and business vocabularies should all be defined and used as baseline for conformance assessments.

SECTION 6 ORGANIZATIONAL MATURITY LEVELS

6.1 Introduction

The maturity of an organization is related to the data quality framework elements described in previous sections. The data quality maturity levels described below provide a means to assess organizational aspects of data quality. Five levels are described, ranging from the initial level, where all data quality activities are *ad hoc* and performed reactively, to the optimized level, where all data quality issues are analysed and remedies implemented proactively, before analytics or operations are affected.

6.2 Level 1 - initial

At the initial level, the organization lacks data quality governance, processes, tools, and metadata for data quality performance and measurement. Although one-off data quality activities may be performed, coordination and systematic learning is not in place, and no best practices are used or established. Furthermore, data quality roles and responsibilities are not defined. There is no recognized link between data quality issues, business impact, and operational risk.

Management perspective on data

Data is managed as a requirement for the implementation of projects.

Pragmatic data quality

Data quality is adequate for local and dedicated usage or in data silos, but data reuse is difficult, the quality is not documented, and data analytics and automation are seldom very useful.

6.3 Level 2 - repeatable

At the repeatable level, the organization has some basic data quality governance, processes, and tools for data quality performance and measurement. Some best practices for methods and processes are used, and learning is taking place. There is increased awareness of the need for technologies and tools to support data quality processes. Some tools are in place, but are not standardized across the enterprise.

Data quality activities are partially or locally coordinated. Data quality roles and responsibilities are partially defined, but are not standardized across the enterprise. The link between data quality issues and business impact is attracting attention and initial work for establishing proactive activities are in place. The organizational ability to respond to data failures is starting to become more streamlined.

Management perspective on data

There is awareness of the importance of managing data as a critical infrastructure asset.

Pragmatic data quality

Data value chains may work under specific and limited conditions, but reuse of data is difficult outside the original designed data flow. Operations with limited/low risks can, with some limitations, benefit from data analytics and automation.

6.4 Level 3 - defined

At the defined level, the organization has data quality governance, processes, and tools in place for data quality performance and measurement. Best practices for methods and processes are used, and learning is taking place. Technologies and tools to support data quality processes are in place and standardized across the enterprise.

Data quality activities are coordinated, and roles and responsibilities are defined and standardized across the enterprise. The first risk analysis of data quality related to business impact has been performed. The link between data quality issues and business impact has attention and some proactive activities are in place.

<i>Management perspective on data</i>	Data is treated as a critical asset for successful mission performance at the organizational level.
<i>Pragmatic data quality</i>	Data can be reused outside original purpose and designated data flow. Data analytics and process automation can work, with caution, in many cases. Functions / operations with moderate risk can benefit from data analytics and automation.

6.5 Level 4 - managed

At the managed level, the link between data quality issues and business impact is described and measured. Relevant proactive activities are in place. Data quality measurements are aligned with requirements and the business impact is known. Risk analysis of data quality measurements is routine, the risk acceptance levels are defined and agreed, and risk management with focus on data quality is in place. Data quality measurements are performed both upstream and downstream in the data value chain. The organizational response to data failures is streamlined and well documented.

<i>Management perspective on data</i>	Data is treated as a source of competitive advantage.
<i>Pragmatic data quality</i>	Data brings considerable value outside their original designated purpose and data flow. Data analytics and process automation provide a competitive advantage. The first fully autonomous systems are in operation. Functions / operations with high risk can benefit from data analytics and automation.

6.6 Level 5 - optimized

At the optimized level, data quality measurements are used as input to innovation and continuously improve data quality and business operations. The effects of data quality improvement activities are measured and the results are used as input to the continuous improvement cycle.

<i>Management perspective on data</i>	Data is considered critical for survival in a dynamic and competitive market.
<i>Pragmatic data quality</i>	Data quality is used to brand the company as a preferred partner. Creation of new knowledge is based on analytics, with high trust in analytical results. Autonomous systems, with high trust and low risk, utilize the data.

SECTION 7 DATA QUALITY PROCESS MATURITY ASSESSMENT

7.1 Introduction

Data quality processes are measured according to data quality maturity levels. The assessment provides a score from 1 to 5, where score 1 indicates a low organizational maturity, with little or no formal support for data quality processes and capabilities. Score 5 indicates high organizational maturity, with well-defined, enterprise-wide data quality capabilities, supporting continuous improvement and strong linkage between business impacts and data quality. The higher levels also focus on information criticality to business operations and dependencies on upstream and downstream information value chains. This typically involves other legal entities that may have other data quality regimes and organizational maturity.

The following sub-sections describe how the result of an evaluation could appear and the criteria for evaluation.

7.2 Maturity score

Maturity heat maps are used to visualize the results from the data quality maturity evaluations. [Figure 7-1](#) shows a matrix of 160 cells, each representing a question from a survey. The survey was structured along maturity levels and the 8 maturity areas. The example in [Figure 7-1](#) indicates that an organization with such a score has more or less fulfilled the requirements for being on the Defined level. Some cells or topics are partly fulfilled. A criticality and risk evaluation of these scores will help with prioritization measures required for improvements.



Figure 7-1 Visualization of maturity score

7.3 Evaluating framework elements

7.3.1 Introduction

All framework elements are evaluated according to maturity level. The evaluation result is input to:

- determine and perform the actions required in order to close the gap between the actual maturity and the target maturity
- data quality risk assessment

The sub-sections below describe the characteristics for each framework element and the corresponding maturity levels. This should be used as high-level guidance for evaluating the maturity level for an organization.

7.3.2 Governance

Initial	At the initial level, there are no formal goals, procedures, or governance for data quality. Responsibilities and activities are not defined, and technology is used in an ad hoc manner to solve any data quality issues.
Repeatable	Best practices are created and published based on previous experiences with resolving data quality issues. There is a bottom-up approach to data quality, where concerned individuals initiate recommendations for data quality techniques, processes, and governance.
Defined	Metrics for evaluation of data quality governance and data quality results are established. Roles, responsibilities, and processes are systematic and used throughout the enterprise.
Managed	Data quality governance measurements and business needs are interlinked and improvements in data quality are prioritized, based on business criticality.
Optimized	At the optimized level, data quality governance is a systematic component of data management and information governance, and tightly aligned with business goals. Continuous improvements are in place and the enterprise uses data quality benchmarking to investigate possibilities for further improvements.

7.3.3 Organization and people

Initial	At the initial level, formal roles and responsibilities are not defined, and there are reactive responses to data quality issues as they occur. There is no data quality culture.
Repeatable	Technical and operational roles and skills are clustered at the department or line of business levels. In some business areas accountabilities for data management are more formalized and described. Some of the capacity needs of the organization are met, and draft descriptions of knowledge and capacity needs are in place. A first version of a roadmap for establishing a data quality culture is in place.
Defined	<p>Formal operational roles and the organizational structure are defined, including a data quality responsible data hygienist-role that ensures that data is measured, cleaned, and made as fit-for-use as possible. The role may be incorporated into formal job titles such as data ingestion engineer, data analyst, and data scientist. Data quality issues are rigorously tracked by designated roles, and the organization includes risk management for data quality.</p> <p>Activities and goals are in place for ensuring a data quality culture. Knowledge requirements and personnel capacity are described, and activities are in place to meet most of the needs. Culture changes and activities are in place according to the road map, and measurements of data quality culture are performed.</p>
Managed	<p>Cross-functional collaboration, including cross-functional business teams, plays an essential role in unlocking the value of data. The data hygienist role ensures that data remain fit for use throughout its entire life cycle, while data stewards are always involved in data quality improvement efforts. Risk assessment of data is performed early.</p> <p>Data quality culture activities are aligned with other cultural activities, and progress is measured as part of the larger picture. Needs for, and access to, knowledge and capacity are managed, and predictions of future needs are in place. Formalized knowledge is built into algorithms and automation.</p>

Optimized	<p>At the optimized level, innovation becomes key in maintaining the vision of improving data quality and addressing data quality issues. Management and data governance staff keep abreast of important emerging trends in data quality management and adapt accordingly.</p> <p>Data quality culture feedback loops are in place. Proactive and corrective actions related to culture are part of everyday life. Knowledge is captured as part of work and made accessible to co-workers. A high degree of formalized knowledge is built into algorithms and automation. Formalized knowledge is transparent, tested, and verified.</p>
-----------	---

7.3.4 Processes

Initial	<p>At the initial level, all data quality issues are handled in an <i>ad hoc</i> and reactive manner that lacks traceability and focuses on symptoms rather than on the root cause. Processes are often manual and performed repeatedly, with no communication with related processes upstream or downstream of the data quality-related incident.</p>
Repeatable	<p>Simple errors can be tracked down and the root cause can be permanently remedied. Rudimentary process support and audit histories for missing data and format errors are available.</p>
Defined	<p>The data quality processes are defined and documented, and are able to identify and trace more complex data quality issues automatically. The processes are implemented at enterprise level and any necessary mitigating activities are specified and communicated to the responsible unit or person.</p>
Managed	<p>All data quality processes are defined by policies and governance, and are performed in a proactive manner. The proactive nature leverages early detection of data quality issues and good damage control.</p>
Optimized	<p>At the optimized level, the main objective of data quality processes is facilitating the continuous improvement cycle. The cycle could be any plan-do-check-act-based methodology and could be formalized by standards such as ISO 9000. All aspects of the processes must be measured, audited, and communicated.</p>

7.3.5 Requirement definitions

Initial	<p>At the initial level, there are no articulated expectations of data quality. Raw data is fed into the system without considering any potential data quality issues.</p>
Repeatable	<p>Basic syntactical data quality issues are discovered and reported. Examples of issues that may be identified are values out of range, invalid data formats, completeness of tuples in datasets, etc. The requirements are not formally stated but can be inferred from domains and contexts, and are often based on profile analytics.</p>
Defined	<p>Data requirement definitions are clearly stated and formalized. Data quality rules for syntactic and semantic compliance are documented and supported by methods and technologies. The requirements are clearly linked to business impacts.</p>
Managed	<p>Conformance with the data requirement definitions is assessed at regular intervals as an integral part of the data ingestion process. The requirements are defined as a component of the data quality policies and governance. The business impact is used as a guide for prioritizing activities to ensure conformance with data definition expectations.</p>

Optimized	At the optimized level, the data requirement definitions are linked to industry improvement targets as an integral part of a continuous improvement cycle. Baselines are defined and the effects of improvement are measured relative to baseline. Enterprise goals are used to modify requirement definitions, and roles are defined to assign responsibility for tracking and collecting changes to requirements.
-----------	---

7.3.6 Metrics and dimensions

Initial	At the initial level, there are no incentives and no defined metrics for measuring data quality. Data quality issues are not categorized according to their general nature and a coherent framework to classify data quality issues is not in place.
Repeatable	Data quality issues are recognized and classified according to data quality dimensions, such as consistency, integrity, accuracy, and completeness. Data quality rules are defined and data value conformance is measured according to defined dimensions.
Defined	Data quality metrics and dimensions are clearly defined, and requirements for syntactic and semantic data quality are specified. Data quality rules are defined and verified both for syntactic and semantic data quality and pragmatic data quality is validated; the results are communicated by data quality reports.
Managed	Data quality metrics are in line with enterprise policies and governance, and the effects of non-conformance are linked to business impacts. Roles are established for responsibility in responding to data quality issues and maintaining definitions of metrics and dimensions.
Optimized	At the optimized level, the data quality metrics and dimensions are linked to SLA specifying threshold values for relevant mitigating actions. Definition and maintenance of data quality metrics are integral parts of IT-system development, ensuring a proactive response to data quality issues.

7.3.7 Process efficiency

Initial	At the initial level, the organization works reactively, using system and operational failures caused by data quality as triggers for data quality activities. Data quality is not monitored.
Repeatable	Generic measures of data quality are performed by profiling, and some areas of impact are defined and linked to the profiling measures. Reengineering is deployed to provide input to more business-related and specific data quality measurements.
Defined	The business impact of data quality issues is analysed and measurement devices are in place to detect data quality issues well in advance of resulting failures. Any data quality issues discovered, and the relevant resolutions, are tracked in a systematic manner. Data quality technology tools for measurements are established.
Managed	Data quality measurements are published as a part of corporate performance reports and displayed in data quality dashboards showing DQI, status, and trends. The reports are communicated and accessible to relevant stakeholders, with focus on any adverse business impacts.
Optimized	At the optimized level, data quality measurements are dynamic and can be updated by business users responding to policy and governance modifications. The measurements are used actively to suggest data quality activities, ensuring support for continuous improvement.

7.3.8 Architecture, tools and technologies

Initial	At the initial level, there are no formal technologies available for performing data quality-related tasks. Tools are developed internally to solve specific issues for particular datasets. Coherent or scalable enterprise architecture is not implemented.
Repeatable	Tools supporting data quality activities, such as profiling, parsing, cleansing, reengineering, decoupling, and entity identification, are available and partially implemented at the repeatable level. The technology is used on an individual basis and does not follow any best practices.
Defined	Best practices for using data quality technology are in place and communicated. The tools support validation of business rules and conformance with standards. The technology is integrated and executed as a part of the IT infrastructure.
Managed	The data quality technology follows defined guidelines, architecture, goals, and governance, and seeks a high degree of automation in correcting data according to business rules and reference data. Data quality analytics can be performed at several steps upstream and downstream in the value chains. Analytics can be performed in real time on, for example, streaming data. The assessment results are visualized and communicated by enterprise-wide dashboards and reports.
Optimized	At the optimized level, data quality tools are used as an integral part of the continuous improvement cycle, providing business users with a low threshold for adjusting data quality rules and assessing the impacts on performance. Change management is used to track any changes to the rule-set. Formalized knowledge and algorithms are transparent, tested, and verified.

7.3.9 Data standards

Initial	At the initial level, standards for metadata definitions, interoperability, quality, security, or other data quality-related processes, techniques, protocols, or architectural elements are not defined. There could be inconsistent data values internally within datasets, as well as across related datasets.
Repeatable	Relevant data standards are identified and guidelines for implementation of the standards are developed at the repeatable level. The requirements for metadata are specified for the identification of datasets, as outlined in the Dublin Core Metadata Initiative or equivalent.
Defined	The standards identified at the repeatable level are fully implemented at the defined level to provide consistency and traceability. Standards for enterprise-internal interoperability are in place.
Managed	All standards and business vocabularies are governed at the enterprise level, and all requirements are communicated to data owners. The overall responsibility for maintenance and conformance with standards is clearly defined within the enterprise and is implemented as an integral part of the IT infrastructure. Standards for data value chain and life cycle management are in place.
Optimized	At the optimized level, all data defined by standards and related processes are handled automatically and supported by policies and governance. Opportunities for improvement by implementing standards are integrated into the governance and are supported by a global taxonomy for data standards.

SECTION 8 DATA QUALITY RISK ASSESSMENT

8.1 General

The consequences and business impacts of data quality issues on different uses and contexts of data should be evaluated by common risk management frameworks, such as those detailed in ISO 31000. The risks associated with data quality should be identified and ranked according to probability and consequences. This section gives a brief outline of how risk management methods can be applied to understand how data quality issues may impact areas such as analytics, operations, automation, integrations and value chains, reporting, and decisions. At the highest level, the business impact is divided into operations, environment, and safety. The main goal of risk analysis is to decide upon, and prioritize, effective actions to mitigate data quality issues and to measure the effectiveness of any actions implemented. As [Figure 8-1](#) illustrates, the process of risk assessment contains the following sequence of tasks:

- define scope and scales for the risk assessment, describing data usage and the datasets to be used
- perform risk identification
- rank risk according to criticality of consequences
- identify, evaluate, and prioritize responses and mitigating actions
- monitor risk scope, changes, incidents, and assessment process continuously
- review to improve the efficacy of the process and repeat based on needs or review criteria.

Scope design is crucial. A narrow scope may result in assessment fragmentation and major threats or vulnerabilities may be outside the defined scope. This may result in a situation where major risks fall outside the scope, overlapping or similar risks may occur in disparate scopes with slightly different wordings, or a major risk may be listed as several non-critical risks.



Figure 8-1 Risk management process

The risks associated with data quality should be identified and ranked according to probability and consequence of potential unwanted data quality related incidents. A risk approach to understand how data quality issues may impact areas such as analytics, operations, automation, integrations and value chains, reporting, and decisions. At the highest level, the business impact is divided into *operations*, *environment*, and *safety*. The main goal of the risk analysis is to decide upon and prioritize effective actions to mitigate data quality issues. In addition, repeating the risk analysis effects of the actions implemented. Examples of common data quality risks in sensor data include:

- Unstable data feeds from source X due to sensor errors => analytics on X are not trustworthy.
- Data values from feed Z are not within boundaries => equipment downtime due to operation outside limits.
- Measurements from feed Y / (second) deviate more than expected from feed Z => either feed Y or feed Z are sending error-prone values and automation fails.

In addition to improve existing data, risk also covers other opportunities. These are identified and analysed as part of the risk process. Examples of data-related opportunities can be creation of new datasets, installation of improved sensors, merging external data sources with internal data sources etc.

A data quality flaw in one element of data may be of limited business impact for a particular area of use, but the same flaw may have critical impacts on other areas of use. [Figure 8-2](#) shows a scenario in which data element 1 is critical for analytics, but has no impact on operations. Thus, different data sources and relevant data quality issues must be categorized according to business criticality for the different areas of

use (context). Identification of risks and the subsequent risk assessment and analysis provide input to guide the prioritization of mitigating actions to risks with the potential for high (business) impact. Risk analysis methodologies provide a wide range of tools for this purpose that can be used in all contexts of data quality issues. Risk contexts can be categorized according to use of data (as listed above), and data quality risk sources can be of various categories, such as systems, networks, logging, integrations, models/ algorithms, format and protocol transformations, sensors and sensor networks, metadata, knowledge, software-tools, governance/management, processes, and metrics.

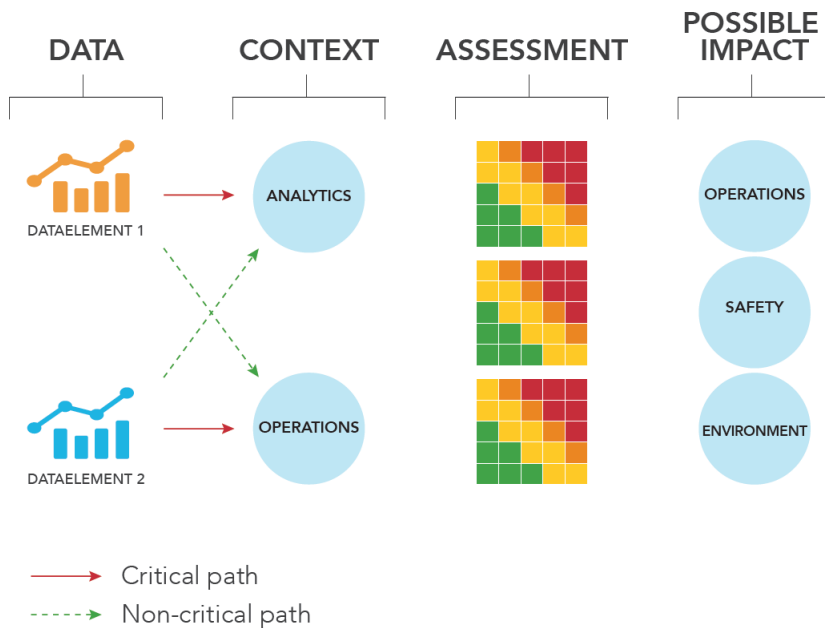


Figure 8-2 Data element risk assessment scenario

Different types of data quality issues impact analytic algorithms and automated processes differently. Currently this relationship or field of knowledge is not thoroughly studied and understood. As knowledge of this relationship increases, DNV GL will be able to perform improved data quality risk assessment.

Figure 8-3 shows a bowtie representation for some typical data quality metrics. The blue ovals on the left-hand side of Figure 8-3 represent the metrics that eventually result in the data quality-related incident and the ovals on the right-hand side of Figure 8-3 represent consequences of the incident. In this case these are categorized as operations, environment, or safety. Possible proactive barriers are shown to the left of the data quality-related incident (in the central yellow circle) and reactive barriers on the right of the incident. Organizations with high data quality maturity levels will have proactive barriers in place, whereas organizations at lower levels tend to apply the reactive measures, after the data quality-related incident has happened. Other bowtie models could be created to represent other levels or contexts of data quality management. This risk assessment enables organizations and actors in value chains to perform risk-based data quality improvements, providing the opportunity for continuous improvement cycles.

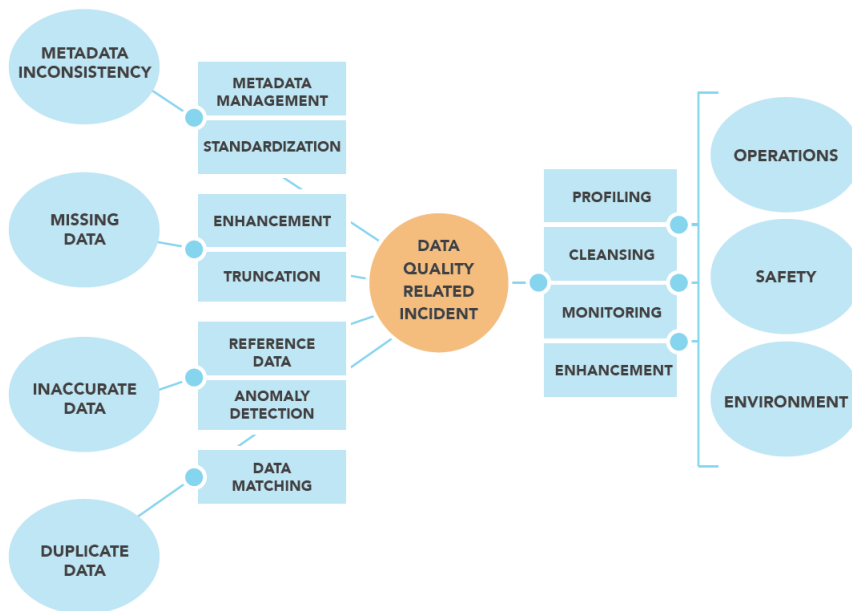


Figure 8-3 Data quality bowtie model

Potential data quality related incidents will be further evaluated in a risk matrix. The DQI described previously is used as input to calculate the probability of potential consequences of incidents. The maturity score will give valuable input to both probability, consequence and which proactive and reactive barriers needed. The scale used for consequence or business impact is discussed with each customer, and evaluated on a scale from low to high. [Figure 8-4](#) is an example of risk matrices for a data quality risk with different score on consequence axis.

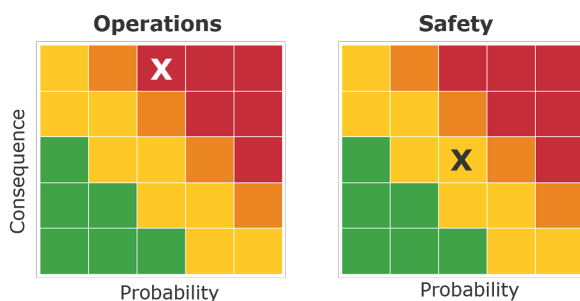



Figure 8-4 Data quality risk matrices

[Figure 8-4](#) show sample risk matrices of data quality risk, e.g., missing values. If a risk is evaluated as high, as in the operations matrix above, mitigating activities should be implemented.

Data quality maturity levels can be assessed in a similar manner for operations, environment, and safety-related consequences. An increase in maturity levels reduces the probability and consequences of risks / data quality-related incidents.

The probability of high consequence data quality-related incidents is believed to be greater for organizations with low levels of data quality maturity. This may often be true, but detailed assessments should be performed for individual cases. Organizations at maturity level 3 or above, can identify high-risk elements



(high probability and high consequence), and can implement mitigation measures for both prevention and correction. Organizations with lower levels of maturity will be unable to rank and address data quality issues in a structured and proactive manner.

The outcome of the risk assessment will be a matrix that indicates whether operations are performed within the chosen limits of risk tolerance. If not, improvements must be performed. In order to choose the best combination of improvement measures, we suggest that an organizational maturity assessment is performed, focusing on data quality processes and capabilities as defined in this RP.

For a comprehensive guide to risk analysis, please refer to ISO 31000 *Risk management* /4/and DNV GL services offered on this topic.

SECTION 9 SECURITY

9.1 General

Information security is an integral part of data quality and data quality maturity. Ad hoc activities can jeopardize integrity, analysis can reveal confidential data, and furthermore, data ownership and governance are necessary to provide the appropriate levels of access control for authorization and authentication. Information security is usually defined by three terms: confidentiality, integrity, and availability. It is important to note that the term integrity has different meanings for information security and data quality. In information security, integrity denotes the unauthorized and undetected alteration of data, whereas in data quality integrity describes references internally or between datasets. An information security integrity breach could therefore result in incomplete and inaccurate data, which are important data quality metrics. [Table 9-1](#) maps information security elements to data quality elements.

Table 9-1 Information security versus data quality

<i>Information security elements</i>	<i>Data quality elements</i>
Confidentiality (sensitivity, access control)	Data governance, policies, and responsibility matrices to ensure data ownership and correct authentication and authorization of users.
Integrity (unauthorized alteration)	Data quality measurements and processes to detect incomplete or inaccurate data.
Availability (denial of service)	Responsiveness, user satisfaction, user feedback loop, accessibility.

Low levels of data quality maturity may adversely affect information security and any data governed by laws and regulations needs to demonstrate some level of data quality governance, policies, and process control in order to comply with regulations. In contrast, low information security could result in incomplete and inaccurate data, as well as reduce general accessibility to the data. Data quality measurements, audit logging, and change management are important data quality activities that are useful tools for monitoring and addressing information security issues. Thus, information security will improve as data quality improves. It is important to keep the relationship between information security and data quality in mind; however, information security must also be addressed as a separate issue, considering the suitability of standards such as ISO 27000 or other required measures. Information security is extensively covered in the literature and further references should be consulted to cover this topic fully.

We recommend that data quality is considered separately from information security. However, the relationships between the two are of great importance. [Figure 9-1](#) shows the relationships between information security elements and data quality elements in a data exchange scenario. The data is transferred from a source system to a target system; the source system must ensure that sensitive information is not uploaded to the target system without the required authorization and security capabilities. In addition, the target system needs to verify and establish SLA covering: (i) data quality requirements, and (ii) security requirements, such as the integrity and availability of the source system.

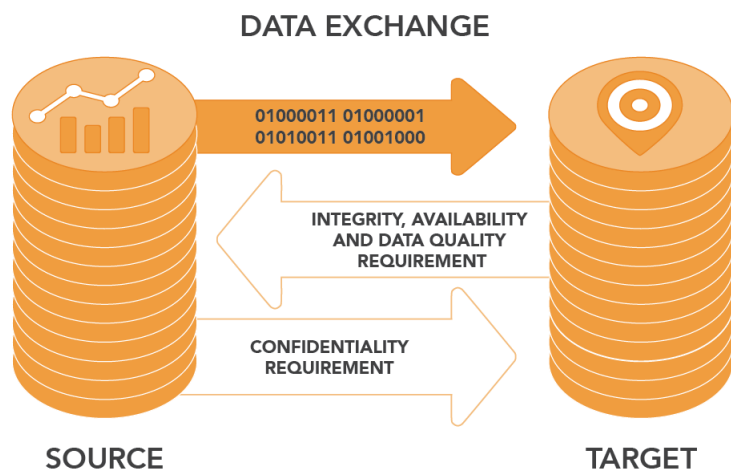


Figure 9-1 Direction of information security and data quality requirements in data flows

Although data quality is considered separately from information security in [Figure 9-1](#), the previously described relationships remain valid.

SECTION 10 REFERENCES

10.1 List of references

- /1/ ISO 8000-8 Information and data quality: Concepts and measuring
- /2/ ISO 9000 Quality management
- /3/ ISO/IEC 27000 Information security management systems
- /4/ ISO 31000 Risk management
- /5/ Data On The Web Best Practices, W3C <http://www.w3.org/TR/dwbp/>
- /6/ Data Management Maturity Model. CMMI institute 2014
- /7/ The Practitioner's Guide to Data Quality Improvements, by David Loshin. MK OMG press 2011
- /8/ ISO 55000 Asset management
- /9/ Journey to Data Quality. Yang et al. MIT Press 2006
- /10/ ISO/IEC/IEEE 15288 Systems and software engineering - System life cycle processes. 2015
- /11/ BS ISO/IEC 25010:2011 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models
- /12/ The Effect of Data Quality on Data Mining - Improving Prediction Accuracy by Generic Data Cleansing, Stang & al, Proceedings of the 14th International Conference on Information Quality, 2009
- /13/ The Effects and Interactions of Data Quality and Problem Complexity on Data Mining, Blake & al, Proceedings of the 13th International Conference on Information Quality, 2008
- /14/ [DNVGL-RP-0496](#) Cyber security resilience management for ships and mobile offshore units in operation
- /15/ The Data Management Body of Knowledge (DAMA-DMBOK Guide) First edition 2010
- /16/ Data Quality Assessment - For Sensor Systems and Time Series Data. DNV GL Technical Report 2017-0058
- /17/ Understanding sensor systems reliability. DNV GL position paper 2-2016.
- /18/ Datiris profiler software, <http://www.datiris.com/metrics.html>
- /19/ A data quality framework applied to e-government metadata. Per Myrseth, Jørgen Stang and Vibeke Dalberg. 2011 International Conference on E-Business and E-Government (ICEE), Shanghai
- /20/ ISO/IEC 11179 Metadata registry
- /21/ Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance' 26 Sep 2013, by David Plotkin

CHANGES – HISTORIC

There are currently no historical changes for this document.

DNV GL

Driven by our purpose of safeguarding life, property and the environment, DNV GL enables organizations to advance the safety and sustainability of their business. We provide classification and technical assurance along with software and independent expert advisory services to the maritime, oil and gas, and energy industries. We also provide certification services to customers across a wide range of industries. Operating in more than 100 countries, our 16 000 professionals are dedicated to helping our customers make the world safer, smarter and greener.

SAFER, SMARTER, GREENER