ABS ADVISORY ON DATA QUALITY FOR MARINE AND OFFSHORE APPLICATION - TIME SERIES DATA

ABS

OUR MISSION

The mission of ABS is to serve the public interest as well as the needs of our members and clients by promoting the security of life and property and preserving the natural environment.

HEALTH, SAFETY, QUALITY & ENVIRONMENTAL POLICY

We will respond to the needs of our members and clients and the public by delivering quality service in support of our Mission that provides for the safety of life and property and the preservation of the marine environment.

We are committed to continually improving the effectiveness of our HSQE performance and management system with the goal of preventing injury, ill health and pollution.

We will comply with all applicable legal requirements as well as any additional requirements ABS subscribes to which relate to HSQE aspects, objectives and targets.

TABLE OF CONTENTS

TABLES

FIGURES

1 INTRODUCTION

1.1 BACKGROUND

Increasingly, modern marine vessels and offshore units are being equipped with diverse functionalities for structural and machinery health monitoring, efficiency monitoring, and operational performance management and optimization. These functions collect data through sensors and onboard instrumentation and analyze that data to provide health and condition awareness, operational and crew assistance, and operational optimization. High quality data is essential in these applications for ensuring accuracy and confidence in analytical results and decision making. Data quality assessment, monitoring, and control are key elements in the data flow associated with such data-centric functions as described in the ABS *Guidance Notes for Smart Function Implementation*, illustrated in Figure 1.

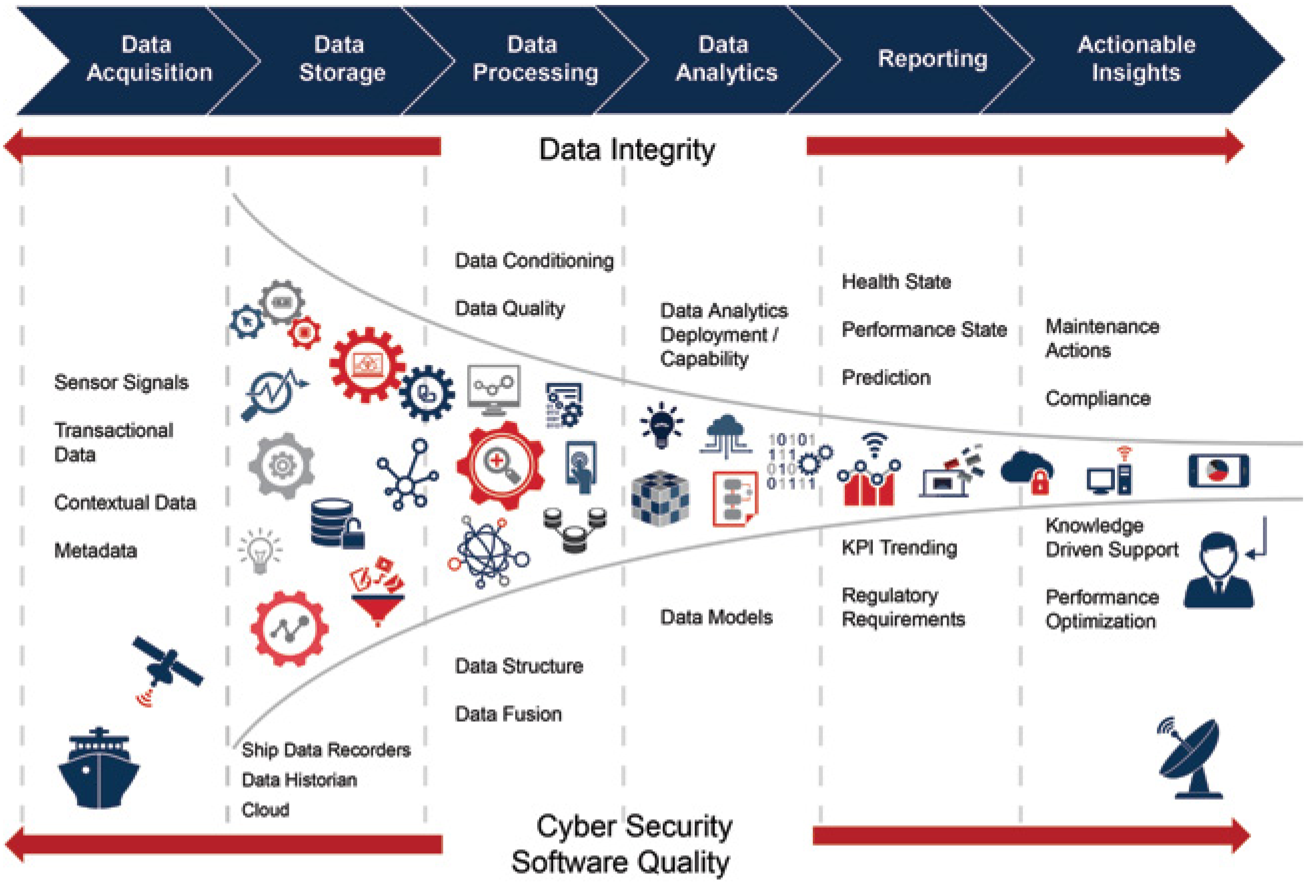


Figure 1: Data Flow Associated with Data-Centric Applications

1.2 GOALS AND SCOPE

This Advisory provides an overview of the relevant standards and industry best practices and offers general guidance and recommendations on data quality assessment, monitoring and control as applied to marine and offshore applications.

Four areas are identified as the focus of this Advisory:

- Overview of international data quality assessment standards;
- Data quality assessment approaches;
- Recommended framework for data quality assessment, monitoring and control;
- Review of typical data quality capabilities available in commercial products.

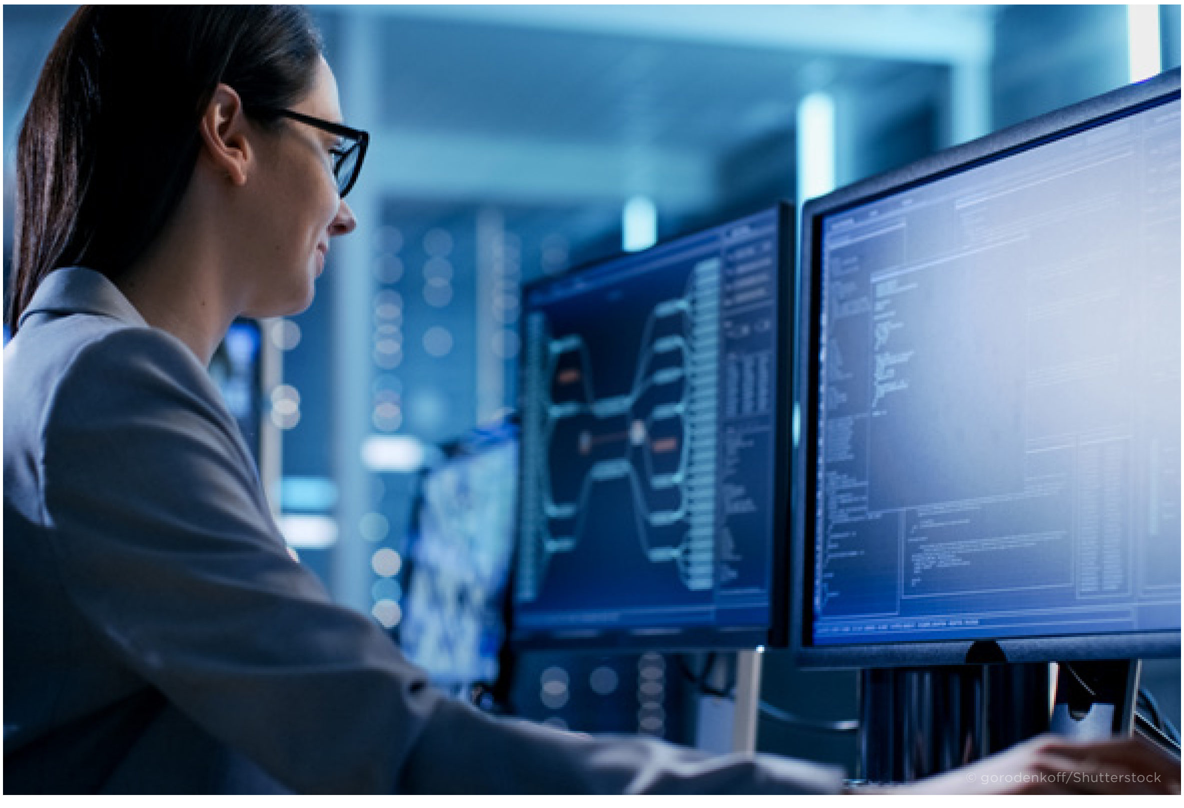
There are many applications for stringent data quality requirements in commercial applications. Data quality criteria is often applied in areas such as business intelligence, customer relations, and numerous applications in the medical and finance industries. This Advisory focuses on marine and offshore applications and looks primarily at how to apply data quality standards to time-series data generated by onboard sensors, measurement instruments, and automation systems (referred to as Internet of Things, or IoT). Realizing that transactional data is commonly leveraged for data-driven marine and offshore applications and decision support, the data quality assessment framework, relevant data quality dimensions and metrics presented in this document can also be applicable to transactional data.

Section 2 reviews the international data quality standards relevant to marine and offshore applications. Section 3 details the recommended data quality assessment approaches and the application of a set of data quality tiers suitable for marine IoT data, which align with the data quality categories specified in the international standards. Section 4 provides a generic data quality assessment, monitoring and control procedure which can be used as a data quality continuous improvement framework for marine and offshore applications. Section 5 discusses the main functionalities and features of several data quality tools commercially available in the market today. Section 6 describes the ABS role in data quality verification and validation. Section 7 summarizes the Advisory.

1.3 CHALLENGES

The main challenges that face the maritime industry with regards to data quality for IoT data can be summarized as:

- Collecting quality data is a challenge in the marine and offshore environments: Marine and offshore environments and operations have unique variables such as noise, dust, temperature, humidity, electronic and magnetic interference, and location since they are typically far from land-based infrastructures. This has a significant impact on the performance, reliability, and longevity of the sensors, cables, and data communication and storage devices. Data collected, transmitted, and stored in these severe environments and operations may potentially have significant quality issues, such as data loss, invalid values, transmission delay, incorrect timestamp order, etc.
- Data quality for IoT data is highly dependent on the data processing quality: Data acquisition and pre-processing occurs when data is collected and converted into a desired form for future analysis. Data integration and transformation (mapping) consolidates and fuses data from different data sources and ensures this data is compatible with the structure of the target data usage. Marine IoT data is typically collected through diverse onboard sensors and data sources with various hardware and software specifications, data format and structures, time intervals, and data definitions, which increases the difficulty of effectively integrating, cleansing and mapping the data.



- Data quality in marine and offshore applications requires domain support: Data quality is not only an IT/data science task. The commercially available automated data profiling tools and software can only help to discover generic data quality problems such as data type mismatch. Data quality assessment, monitoring and control should be considered in the context of the data applications and the overall implementation objectives including business goals. The involvement of business and technical subject matter experts is critical in the process to successfully design, develop and improve the data quality validation rules and means for measuring, monitoring and improving data quality. The usage of domain knowledge is mainly needed to define data quality rules or violations. Thereof, using contextual knowledge from domain, these would vary from one application to another and requires intricate knowledge of the particular domain.
- Data quality awareness and knowledge in the marine and offshore industries is not fully mature: Improving data quality requires a cultural shift within organizations. To improve maturity, managerial accountabilities should be established in organizations to build a culture that values quality data. Establishing a process management cycle (i.e. "define-measure-analyze-improve-control") is beneficial to data quality awareness and improvement in marine and offshore industries.

  The key elements of the process management cycle for improving data quality include:

  - Identifying the critical data issues and business rules,
  - Assessing data against expectations,
  - Identifying and prioritizing opportunities for improvement based on the findings and feedback from stakeholders (data stewards, business and technical subject matter experts, and data consumers),
  - Measuring, monitoring and reporting on data quality,
  - Improving data quality by incorporating incremental changes in the business cycle, such as installing an improved data collection system,
  - Integrating data quality controls into business and technical processes to prevent issues from recurring.

This Advisory is designed to help those who are facing the above challenges and to assist in driving the industry towards better and more reliable data-centric applications.

2 THE ISO DATA QUALITY STANDARD

2.1 ISO 8000

ISO 8000 is the international standard for data quality. ISO asserts:

"The ability to create, collect, store, maintain, transfer, process and present data and to support business processes in a timely and cost-effective manner requires both an understanding of the characteristics of the data that determine its quality, and an ability to measure, manage and report on data quality."

ISO 8000 denotes methods to manage, measure and improve data quality for specific types of data including, but not limited to, master data, transaction data and product data. ISO 8000 can be used independently or in conjunction with quality management systems.

2.2 THE PURPOSE OF ISO 8000

The purpose of ISO 8000 is to make it easier to contract for quality data and to identify companies and software applications that can deliver quality data. In doing so, it is intended that ISO 8000 is to help organizations define what is and is not quality data, enable them to ask for quality data from the supplier using standard conventions, and verify that they have received quality data using those same standards.

2.3 STRUCTURE OF ISO 8000

ISO 8000 defines relevant characteristics of data quality, specifies requirements applicable to those characteristics, and provides guidelines for improving data quality. ISO 8000 is applicable within all the stages of the data life cycle. ISO 8000 is organized as a series of parts, published separately:

- Parts 1 to 99: General data quality
- Parts 100 to 199: Master data quality
- Parts 200 to 299: Transaction data quality
- Parts 300 to 399: Product data quality

The relevant ISO development efforts (including published standards, and standards under development) can be tracked from the Standards Catalogue under ISO Technical Committee 184 Subcommittee 4 (ISO TC184/SC 4)[1]. The committee's mission is to develop standards for the exchange of complex data in an application-neutral form.

2.4 PART 8 OF ISO 8000

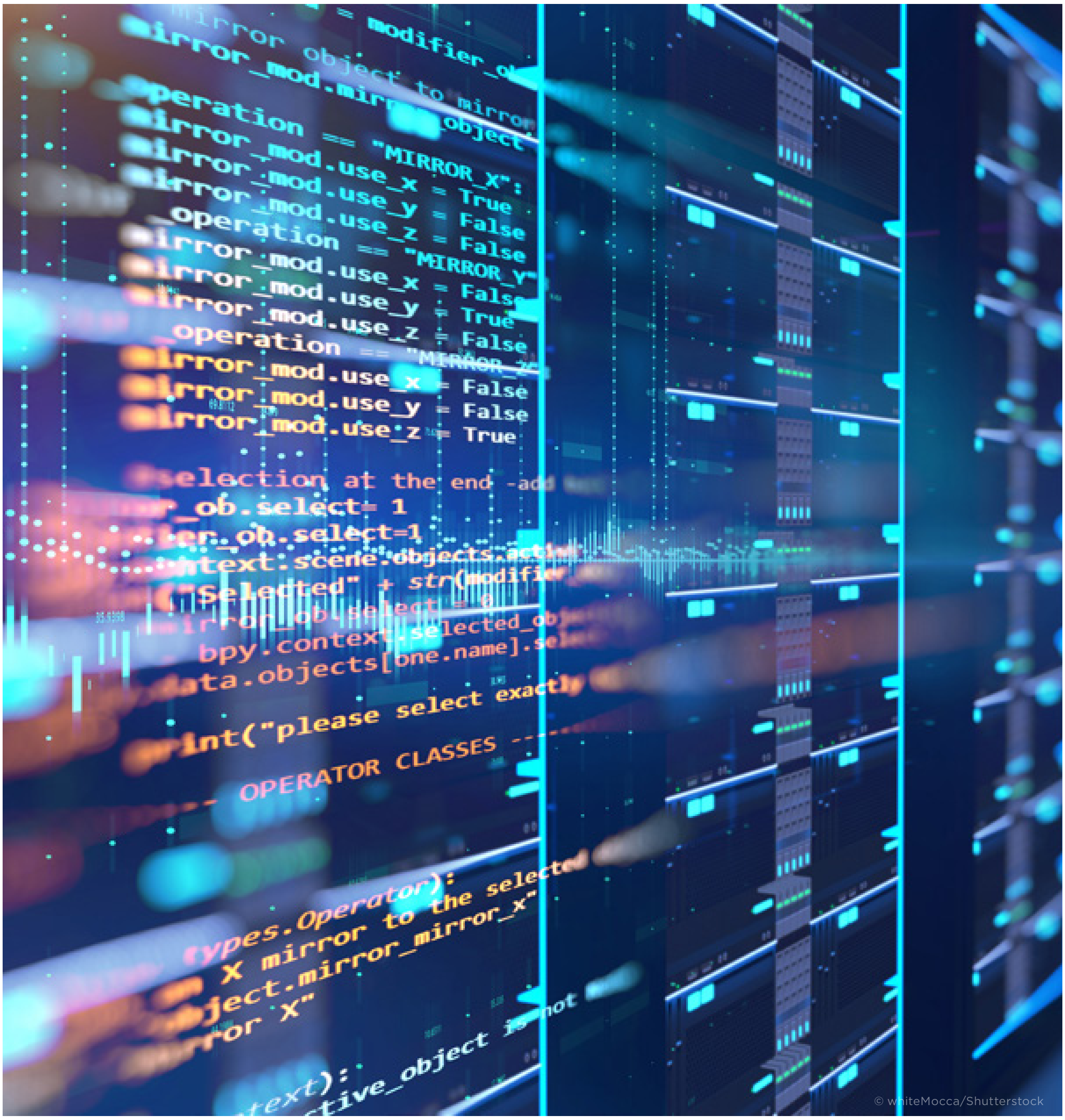
Part 8 of ISO 8000 (ISO 8000-8:2015, November 2015) provides fundamental concepts to plan and perform data quality measurements. Its application is independent of status of organization, type of information or data, hardware storage medium, software, information security and information life cycle stage. The main purpose of ISO 8000-8 is to provide a foundation for measuring data quality according to the following three categories:

- Syntactic quality – the degree to which data conforms to its specified syntax, i.e. requirements stated by the metadata;
- Semantic quality – the degree to which data corresponds to what it represents;
- Pragmatic quality – the degree to which data is found suitable and worthwhile for a particular purpose.

Measuring syntactic and semantic quality is performed through a verification process, while measuring pragmatic quality is performed through a validation process.

ISO 8000-8 also provides the prerequisites for measuring data quality, and it specifically describes requirements along with a set of data quality rules/dimensions for data quality verification and validation pertinent to each of the three data quality categories. Meanwhile, some narrative examples are provided to describe violations of the rules.

[1] https://www.iso.org/committee/54110/x/catalogue/p/1/u/0/w/0/d/0



ISO 8000-8 does not:

i) Provide visualized interrelationships among the three data quality categories. Although they are separated for analytical convenience, they are closely interrelated and build on each other.

ii) Establish the thresholds and acceptance criteria used for measuring data quality. They are dependent of the intended data application. Data quality analysts or data consumers need to set the acceptance criteria and weighting factors for each applied data quality dimension and metric prior to measuring data quality.

iii) Map the concepts of data quality measurements into a data quality assessment, monitoring and control system. It is required to develop procedures for continuous measurement of data quality and improve the process of data quality management.

Sections 3 and 4 of this Advisory address the limitations of ISO 8000-8. A procedure for data quality assessment, monitoring and control is defined to leverage the three data quality categories into a data quality management and continuous improvement system.

3 DATA QUALITY ASSESSMENT TIERS

A set of data quality dimensions frames the business requirements for data quality. For various data-centric applications and their implementation goals in marine and offshore applications, the needed function's performance and reliability vary according to the function's role in decision-making. The requirements on the data quality are different for various applications. To adapt the different levels of the data quality assessment for IoT data, a tiered data quality assessment framework can be established which aligns with the three data quality categories specified in ISO 8000-8. The interrelationship among the three assessment tiers is illustrated in Figure 2. Typically, the Tier 3 application level assessment is performed on the basis of the Tier 1 generic level assessment and the Tier 2 sensor/equipment level assessment.

- Tier 1: Generic Level - Syntactic Quality Verification: Verifies the degree to which data conforms to its specific syntax. The assessment goal is the consistency where data values for particular data elements in the database/storage must conform to the pre-defined data content and structure as represented by the metadata (e.g., data hierarchy, data field type, data format, value domain, precision, etc.). The Generic level quality check can be applied to any IoT data in marine and offshore applications.
- Tier 2: Sensor / Equipment Level - Semantic Quality Verification: Verifies the degree to which the data corresponds to what it represents (e.g. how accurately the measurement reflects the sensed physical value). The assessment goals are the data's comprehensiveness and accuracy. The Sensor/Equipment level quality check requires knowledge of the sensor, equipment and system working principles, measured physical and environmental parameters, and operational conditions.
- Tier 3: Application Level - Pragmatic Quality Validation: Validates the degree to which data is found suitable and worthwhile for a particular purpose. The assessment goals are the data's usability and usefulness for a given use. Usability is intended for validating if the data consumers can easily access, effectively retrieve and manipulate the data in the database/storage. Usefulness is intended for validating if the data can support the consumers in data applications for accomplishing tasks. In addition to the commercial and technical domain support, the application-level quality check may require IT and data analytics skills and knowledge. This may include confirming the data is accessible and interpreting unclear physical relationships, health status, and/or complicated physical and operational conditions derived by an employed data analytics model.

TIER 1 GENERIC LEVEL

Syntactic quality verification

- **Data Quality Goals: Consistency**
- Focus on the structure (symbolic forms) of data
- Require data quality metadata (Unique identifiers, data type, format, unit, value range, etc.)
- Identify Syntactic data quality rules, dimensions and metrics

TIER 2 SENSOR/ EQUIPMENT LEVEL

Semantic quality verification

- **Data Quality Goals: Comprehensive and Accuracy**
- Focus on the meaning of data
- Require stakeholders' (business and technical subject matter experts, data users) domain knowledge and experience
- Identify Semantic data quality rules, dimensions and metrics
- Set weighting factors, thresholds, acceptance criteria, score ranges for a particular sensor/ equipment

TIER 3 APPLICATION LEVEL

Pragmatic quality validation

- **Data Quality Goals: Usability and Usefulness**
- Focus on the usage of data
- Require stakeholders' feedbacks
- Identify Pragmatic data quality rules, dimensions and metrics
- Set weighting factors, thresholds, acceptance criteria, score ranges for a particular application

Figure 2: Three Data Quality Assessment Tiers for IoT Applications

The data quality dimensions represent the measurable characteristics of data. The applicable data quality dimensions for typical marine and offshore IoT data are specified in Appendix 1 for each data quality assessment tier. Each data quality dimension consists of a single or a set of data quality metrics, which is an indicator of the data quality performance of data records in a specific aspect. The metrics are typically measured via the percentage of data that conform to certain data quality validation rules. Both data quality dimensions and metrics should be quantitively measurable. A typical hierarchy of data quality assessment results (scores) is presented in Figure 3.

The data quality validation rules defined for employed metrics are constructed based on the knowledge and understanding of the use of the data and potential data quality problems associated with the data for certain data-centric applications.

For a specific data quality tool, the data quality validation rules, metrics and dimensions should be configurable for different data sets and various analytics objectives. Appendix 2 provides examples of data quality validation rules and dimensions targeting potential data quality problems. The methodology and procedure used for determining/selecting data quality requirements is explained in further detail in Section 4.

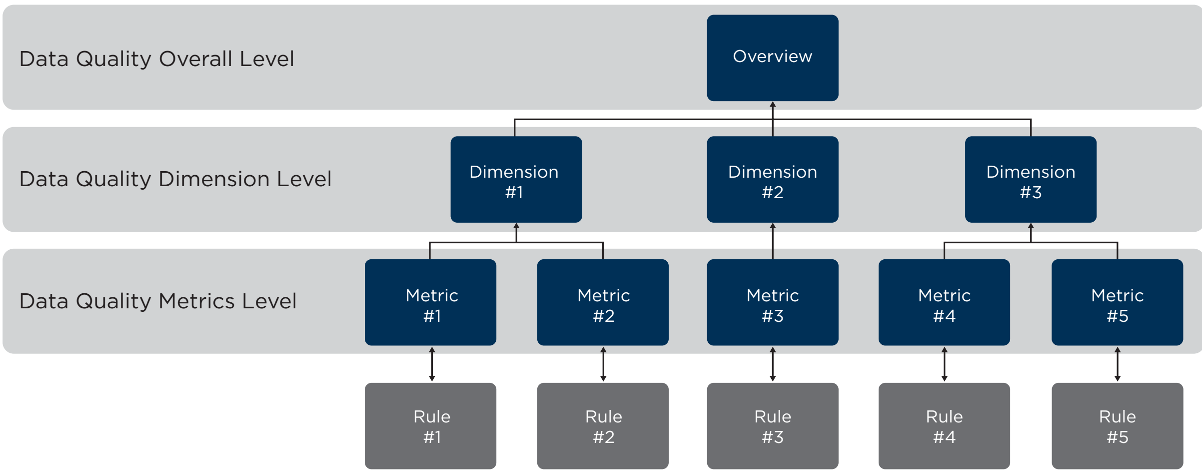


Figure 3: A Hierarchy of Data Quality Results

4 DATA QUALITY ASSESSMENT, MONITORING AND CONTROL PROCEDURE

A systematic approach can greatly assist the planning, design, development, and implementation of the data quality assessment, monitoring and control procedure. For various levels of data application sophistication, different approaches may be adopted. In this advisory, a recommended data quality assessment, monitoring and control procedure (see Figure 4) is established in line with the common quality improvement cycle "plan-do-check-act" (ASQ, 2019).

4.1 KEY STEPS OF DATA QUALITY MANAGEMENT

The recommended data quality assessment, monitoring and control procedure as shown in Figure 4 illustrates an iterative data quality management method. The key steps of which are discussed in the following sections.

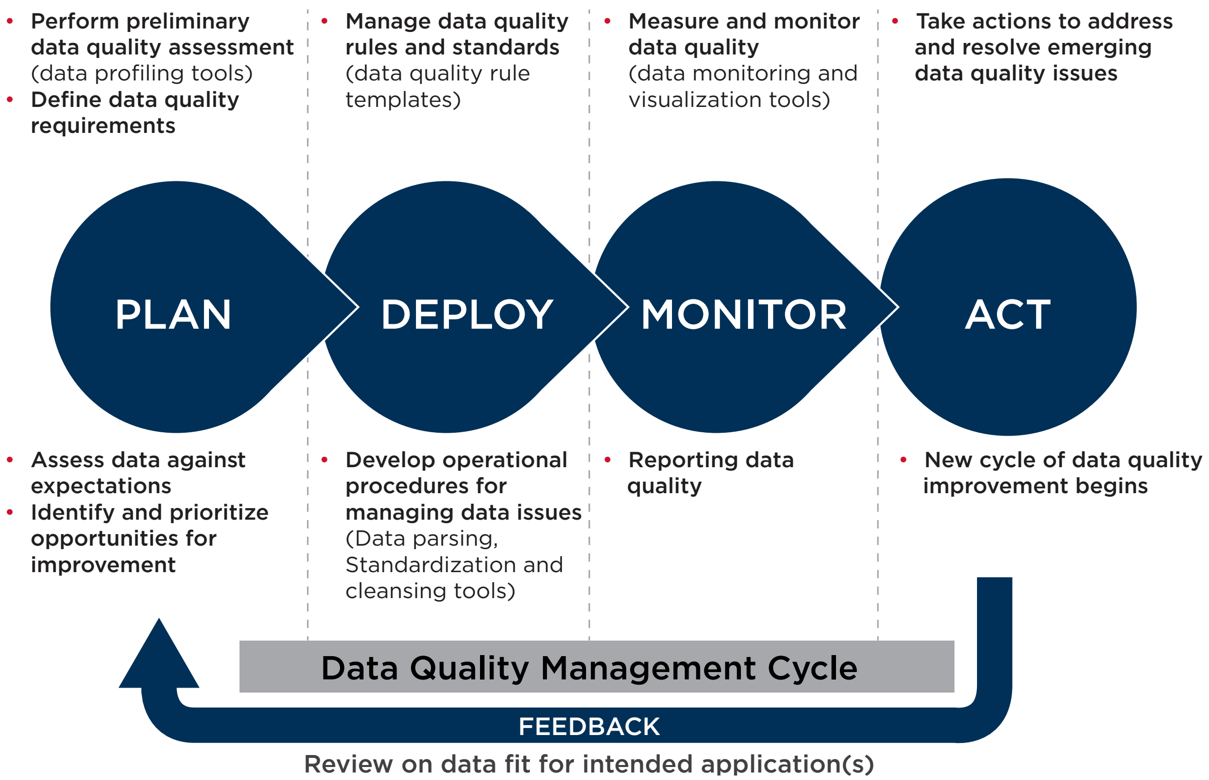


Figure 4: Data Quality Assessment, Monitoring and Control Flowchart

STEP 1 - FORMULATE A BASIC PLAN FOR DATA QUALITY ASSESSMENT AND IMPROVEMENT

At this Plan step, the data quality requirements should be defined based on the identified data quality issues through preliminary data profiling. A basic data quality improvement plan should be prepared to address the identified data quality issues. The main tasks and activities at this step include:

a. Perform preliminary data quality assessment

 i) Select/Identify a data set for initial assessment (e.g. a small or a full-scale data set from the data sets to be assessed or from historical data sets generated by a similar data application) by using classical statistical analysis approaches (e.g. data profiling and similar techniques). The process of profiling and analyzing the data set helps identify the data characteristics (e.g. data type, format, precision, range, etc.), discover potential data quality issues and define the data quality validation rules.

 ii) Identify potential data quality issues/anomalies through the data profiling of the identified data set, as well as the potential issues from data collection, transmission, and storage caused by the both hardware and software, with the help of subject matter experts.

iii) Evaluate potential impacts of the identified data quality issues on the potential data use (e.g. the impacts on data analytics accuracy and reliability and ultimately the overall functionality of the data application), which requires input from stakeholders along the data chain from both technical and business aspects

b. Define data quality requirements – two approaches for identifying data quality requirements are commonly adopted in this step, which are described in detail in Section 4.2

i) Define data quality requirements according to the identified data quality issues and their potential impacts. The requirement definition includes data quality validation rules, measurable metrics and dimensions (Refer to Appendix 2)

ii) Define the thresholds, weighting factors and acceptance level for each data quality assessment level: metrics, dimension and overall consolidation

c. Assess data quality based on the data quality validation rules

i) Test and validate the defined data quality validation rules by applying them against a test data set other than the full-scale data sets used for the initial assessment in a)

ii) Review the defined data quality validation rules with the users to make sure that they understand them

iii) Evaluate data quality levels based on the data quality validation rules (as initially validated per item i above) and the defined acceptability and thresholds (examples are shown in Figure 5 and Figure 6)

iv) Document levels of non-conformance and observed data quality issues

Dimension	Weight	Score %	Acceptance %
Dimension 1	20	95	90
Dimension 2	50	87	90
Dimension 3	20	100	90
Dimension 4	10	50	90
OVERALL	100	87.5	90

Figure 5: Data Quality Dimension Scores and the Overall Data Quality Score (Weighted Average of Dimensional Scores)

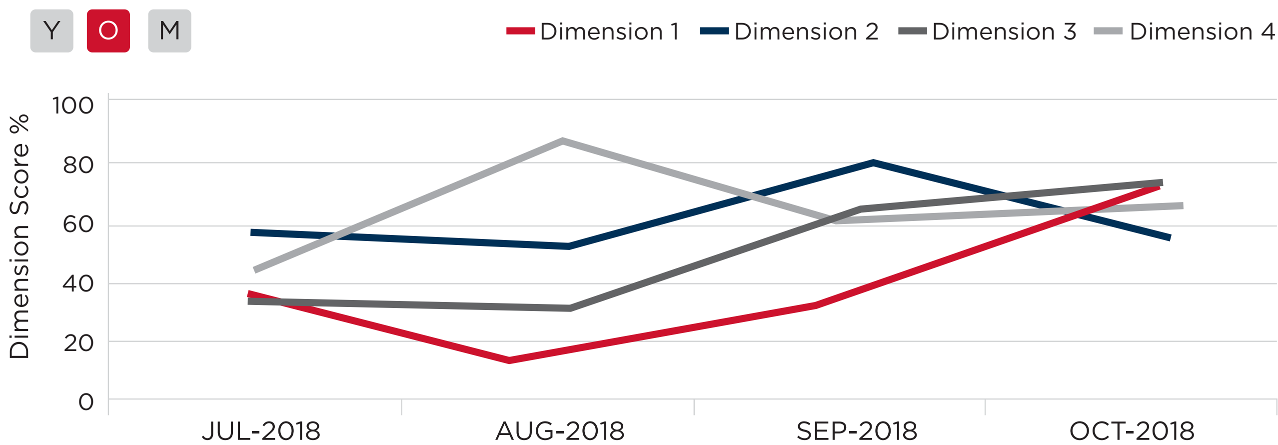


Figure 6: Data Quality Dimension Score Trends

d. Identify and prioritize potential improvements

i) Prioritize criticality the known data quality issues based on the impact on data use and evaluate improvement alternatives to address data quality issues

ii) Prioritize the remediation and improvement efforts, which requires a combination of a full-scale data profiling and inputs from relevant stakeholders

iii) Identify the data quality issues which requires in-depth analysis for the root causes determination and potential improvement alternatives

STEP 2 - DEVELOP AND DEPLOY DATA QUALITY OPERATIONS FOR HANDLING DATA ISSUES

At the Deploy step, the defined data quality requirements from Step 1 should be configured and managed during the operational data quality measurements. After a suitable monitoring period, an operational remediation plan should be developed and implemented for handling the data quality issues.

a. Manage data quality validation rules and standards

i) Rules should be documented with a consistent format with a clear natural language description

ii) Rules should be defined in terms of measurable data quality metrics and dimensions

iii) Rules should be created with consideration of data use. The defined data quality validation rules need to be tested against actual data sets as identified in Subsection 4.1/Step1/c/i). Continuous refinement of the validation rules is recommended throughout the data quality improvement lifecycle

iv) Data consumers and subject matter experts should be involved in defining the data quality validation rules. The defined rules should be confirmed by the data consumers (e.g. who oversee the data use) and subject matter experts

b. Develop operational procedures for handling data issues

i) Diagnose data quality issues

- Review the data quality problem as identified in Subsection 4.1/Step 1/a. and discover the potential root causes of the problem with assistance from the data consumers and subject matter experts

ii) Identify options for addressing data quality issues

- Address non-technical root causes: possibly provide proper training to data handlers; improve the data handling procedure; enhance leadership support; establish clear accountability and ownership
- Address technical root causes: may correct flawed data directly; improve the performance of data collection (e.g. sensors), transmission and storage; modify systems and technical processes to prevent the issue from recurring; continuous monitoring and taking no immediate actions after balancing the impact of the data quality issues versus the cost of the corrective/ improvement actions

iii) Resolve data quality issues

- Perform cost-benefit analysis to compare the potential remediation options as identified in Subsection 4.1/Step2/b/ii). Positive return on investment (ROI) for improvements should be achieved.
- Give advice from the data consumers and subject matter experts to select the best option to resolve the issue:
  - Simple remediation: Fixing and correcting the data directly in records (e.g. data cleansing/ data parsing and formatting)
  - Remediation of root causes: Formulating a long-term improvement plan for strategic changes (e.g., modification of the systems). It focuses on modifying the systems to resolve root causes and putting in place mechanisms to prevent issues in the first place. Prevention is generally more cost saving than correction.
- Develop and implement a remediation plan which intends to re-evaluate the quality level of the remediated data set and to ensure the applied changes do not introduce additional errors, and perform as expected.

STEP 3 - CONTINUOUSLY MEASURE AND MONITOR DATA QUALITY

a. Set a time interval for periodically assessing the data quality against the defined rules

b. Visualize and monitor the data quality results/scores in hierarchical means (e.g. metrics, dimensions and overall levels)

c. Apply a threshold(s) or acceptance criterion for each measurement. The data quality results often reflect the percentage of correct data (passing the validation rule) or the percentage of exceptions (failing the validation rule) depending on the formula used.

   i) Confirm that the data is fit for its intended application (e.g. data analytics) if the data conforms to the defined data quality validation rules

   ii) Notify and alarm data quality issues timely and recommend potential actions according to the developed remediation plan when the data does not conform to the defined data quality validation rules

d. Monitor and trend the data's ongoing conformance with validation rules, and report on all data quality assessment levels

   i) i) Data quality scorecard provides data quality scores in metrics, dimensions, overall level and dashboards related to the execution of data quality validation rules

   ii) Data quality trend shows the data quality changes over time

   iii) Data quality issue management tracks the data quality issue handling according to the remediation plan

STEP 4 - TAKE ACTIONS TO ADDRESS AND RESOLVE EMERGING DATA QUALITY ISSUES

Continuous improvement is achieved by starting a new cycle, which can be seen in Figure 4. New cycles of data quality improvement may restart as

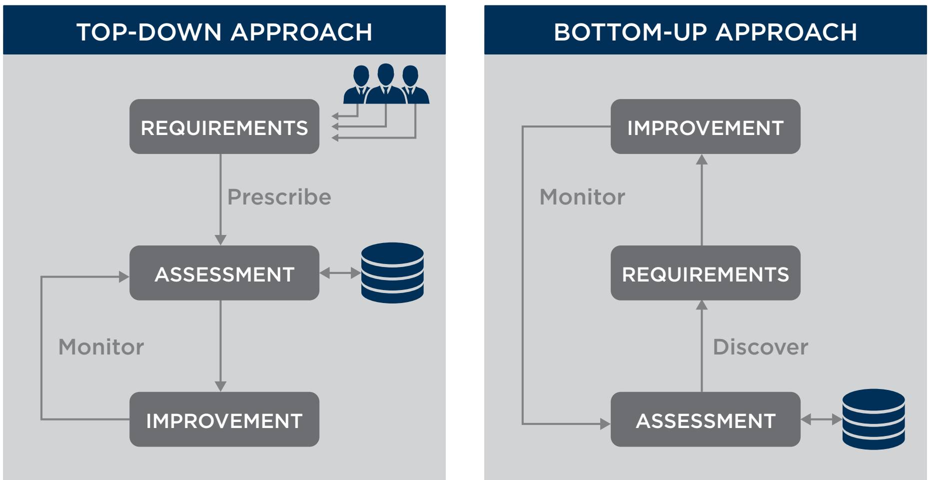
- Existing data quality results fall below acceptance criteria
- New data sets come under investigation
- New data quality requirements are identified for existing data sets

4.2 APPROACHES FOR DATA QUALITY REQUIREMENTS IDENTIFICATION

The data quality requirements considered in this Advisory are applied to validate if the outputs from data processing are fit for data analytics (see Figure 1). The data quality is assessed against a set of initial requirements (ISO/TS 8000-1:2011, 2011). Data quality requirements analysis often includes surveys of data users and subject matter experts to identify data quality issues, which can be used to identify critical data sets, define the appropriate data quality rules, metrics and dimensions, and set quality targets.

Figure 7 compares the two approaches that can be used to define data quality requirements: top-down and bottom-up. (DAMA International, 2010; Zhang, RJ, etc., 2014).

Figure 7: Two Data Quality Requirement Identification Approaches



TOP-DOWN

In the top-down approach, the pre-defined prescriptive requirements of the data quality are initialized by the data consumers who are knowledgeable of the subject data issues associated with its applications and impacts. The requirements are converted into data quality validation rules and organized into metrics and dimensions. The data quality assessment and monitoring are executed based on the defined requirements. This is a data consumer-driven approach, requiring the data consumer to be familiar with the subject data sets and knowledgeable of the features of the intended data application, the critical data dependencies, and the potential data issues that are significant to the success of such applications.

BOTTOM-UP

In the bottom-up approach, the data is initially assessed through data exploration or data profiling to uncover any data anomalies, the requirements on data quality are continuously reviewed and updated according to the potential issues revealed. The data quality is then measured and monitored against the requirements identified from the previous step. If the quality level is below acceptable levels, it should trigger actions to improve the data. The defined requirements must be reviewed and refined accordingly when new data sets come under investigation. This is a data-driven approach, which allows requirements to be dynamically discovered and adapted as the use and understanding of the data set expands, which is suitable for assessing large and unfamiliar data sets.


© Dragon Images/Shutterstock

5 DATA QUALITY TOOLS AND TECHNIQUES

The data quality tool market continues to innovate as more organizations recognize the impact of poor-quality data and seek solutions for improvement. The data quality tools available in the market mainly focus on three data domains: customer data, financial data and product data. The technology leaders currently extend their data services to emerging areas, such as big data, cloud computing, data governance, IoT and machine learning.

The main capability of data quality tools can be sorted into four categories: *Analysis, Cleansing, Enhancement and Monitoring* (DAMA International, 2010; DAMA International, 2017). The available tools on the market are mainly used for data profiling, parsing, transformation and standardization, cleansing, matching, enrichment and monitoring. The capabilities for those tools are briefly introduced below.

DATA PROFILING

Data profiling tools produce data statistics that enable the data analysts to understand the data patterns and data quality issues. Profiling tools have strength particularly in data discovery because of the capability to handle large data sets. The built-in data visualization capability of the profiling tools is beneficial to the process of data discovery. The revealed data quality issues require data consumers and subject matter experts to interpret/identify the root cause and evaluate the potential impact. Table 1 lists five main types of analysis performed by data profiling tools.

Table 1: Types of Analysis Performed by a Data Profiling Tool

Uniqueness Analysis and Completeness Analysis	
Uniqueness	Identifies the percentage of the unique values for individual columns
Completeness	Identifies the percentage of rows that contain actual values for individual columns
Counts of Null	Identifies the number of rows that do not contain values (null) for individual columns
Pattern Analysis	
Data Type	Identifies the syntactic patterns of the data (e.g. the code "W" means a word and "N" means a number) and the total number of each pattern counts for individual columns based on the value contents
Data Format	Identifies the syntactic formats of the data (e.g. the code "L" means a letter and "D" means a digit) and the total number of each format counts for individual columns based on the value contents
Precision	Identifies the largest number of digits in a number for individual columns, which includes digits on both sides of the decimal point (e.g., the precision for "DDD.DD" is 5)
Scale	Identifies the greatest number of digits to the right of the decimal point for individual columns (e.g., the scale for "DDD.DD" is 2)
Min/Max Length	Identifies the overall Min / Max value length for individual columns
Range Analysis	
Min /Max Value	Identifies the overall Min / Max value for individual columns across all data types
Values Distribution Analysis	
Frequency Distribution	Identifies how many times each value in the selected column occurs
Quantile Distribution	Identifies the data values in the selected column that occur at designated intervals in the ordered data set. The first value in the list is at 0% and the last value is at 100%. The median value is at 50%

DATA PARSING AND FORMATTING

Data parsing is the process of deconstructing data into its component parts and formatting the values into consistent layouts based on the user-defined business rules and knowledge bases of values and patterns. Data parsing tools enable the data analyst to define sets of patterns that feed into a rule engine used to distinguish between valid and invalid data values.

If a data record is represented in different formats, a data parsing tool has the built-in ability to extract and rearrange the separate value components into a standard representation to create a valid pattern. When an invalid pattern is detected, actions are triggered to transform the invalid value into a standard format.

DATA TRANSFORMATION AND STANDARDIZATION

A data transformation or standardization tool has the built-in ability to modify data for specific formats, values and layouts by employing industry or local standards, business rules or knowledge bases. Standardization is a special case of transformation that is performed by mapping data from a source pattern into a corresponding target representation. A good standardization tool will be able to parse the different components of a data value and then rearrange, correct or change those components into a target representation.

DATA CLEANSING

Data cleansing is the process of detecting and correcting (or removing) data errors from a record data set or database, and then replacing or modifying data to make it conform to business rules or knowledge bases. Data cleansing tools can help to keep the data clean and consistent, but continuous application is usually costly and may introduce risk. The need for data cleansing should decrease over time. However, in some circumstances, correcting data values may be a cost-effective solution for simple and fast resolving data issues (e.g., a simple remediation plan as identified in the "Deploy" stage of Figure 4).

MATCHING AND LINKAGE

Data matching and linkage is the process of identifying, linking and merging the related data records within or across datasets using a variety of techniques, such as domain rules, algorithms, metadata and machine learning. Data matching tools allow the data analysts to identify duplicates or possible duplicates in large data sets, and then take actions such as merging the identical or similar data elements into one. Identifying similar records within the same data set likely indicates duplicate records and may need de-duping and/or elimination. Identifying similar records in different sets may indicate a link across the data sets, which helps facilitate cleansing, knowledge discovery, and reverse engineering.

Deterministic and probabilistic are the two common types of data matching approaches that help save time and cost for data cleansing and removing duplicates from databases during data processing.

DATA ENRICHMENT

Data enrichment or enhancement is a process of adding new data elements to an existing data set to increase its quality and usability. Data enrichment usually requires integrating externally sourced data (e.g. time/date stamps, reference vocabularies, contextual information, geographic information and demographic information) to improve completeness and add value.

DATA QUALITY RULE TEMPLATES

Data quality rule templates enable the data analysts to understand data use expectations. The templates include the rule descriptions using a consistent format and the rule formulations for data quality assessment at different levels (i.e. three data quality tiers as specified in Section 3) Some common templates can be considered to specify data quality validation rules, such as:

- *Value domain conformance:* Rules that define a data element's assigned value must be within a defined data value domain.
- *Value range conformance:* Rules that define a data element's assigned value must be within a defined numeric, lexicographic (i.e. alphabetical order), or time range.
- *Format conformance:* Rules that define a data element's assigned value must conform to the pre-defined patterns, such as the different ways to specify a timestamp.
- *Record completeness verification:* Rules that define the conditions under which missing values are acceptable or unacceptable.
- *Consistency verification:* Rules that specify certain relationships which need to be maintained between two (or more) data records in the same data source (e.g. the sampling rates of data records in the same time series need to be consistent).
- *Accuracy verification:* Rules that specify the ways to compare a data value against a corresponding value in a system of record or other verified data source (e.g. benchmarking database), and to verify that the values match.
- *Uniqueness verification:* Rules that specify which data elements must have a unique record against the corresponding real-world objects.
- *Timeliness validation:* Rules that indicate the expectations for accessibility and availability of data (e.g. The age of the data is fit to its use).


© Gorodenkoff/Shutterstock

DATA MONITORING

Data monitoring tools provide a user interface to visualize and report results associated with data quality measurement, metrics and activity. The data quality results can be visualized and reported through multiple levels via scorecards and dashboards (drilldown). Data monitoring tools also assist the ongoing assurance and control of data quality to ensure ongoing conformance of data to business rules.

There are commercial products which provide data profiling, parsing, transformation, and cleansing functions with a capability to create data quality rules (validation rules and correction rules). By tracking the number of discovered flaws as a percentage of the size of the entire data set, these tools can provide percentages of conformance to the defined rules at multiple levels (see Figure 3). The next step is to assess whether the level of conformance meets the data use expectations.

The identification of meaningful data quality validation rules and measurable metrics are the key for the downstream data quality assessment in the whole data chain. Domain-specific knowledge can greatly improve this process. The existing commercial data quality solutions do not have the ability to define the domain-specific rules relevant to marine equipment/structure/performance applications. The solutions available in the market are more applicable in the level of generic data quality assessment and data profiling, which can help users identify potential issues and understand the content and structure of the data quickly.

6 ABS' ROLE

As a classification society, ABS has published *Guidance Notes on Smart Function Implementation* (the Smart Guidance Notes) to help guide data-centric marine and offshore applications for structural and machinery health monitoring, asset efficiency monitoring, operational performance management, and crew assistance and augmentation. This Guidance Notes sets an actionable goal-based framework for Smart function implementation along with recommended risk-informed verification and validation principles to provide confidence in the ability to conform to the functional requirements (see Figure 8).

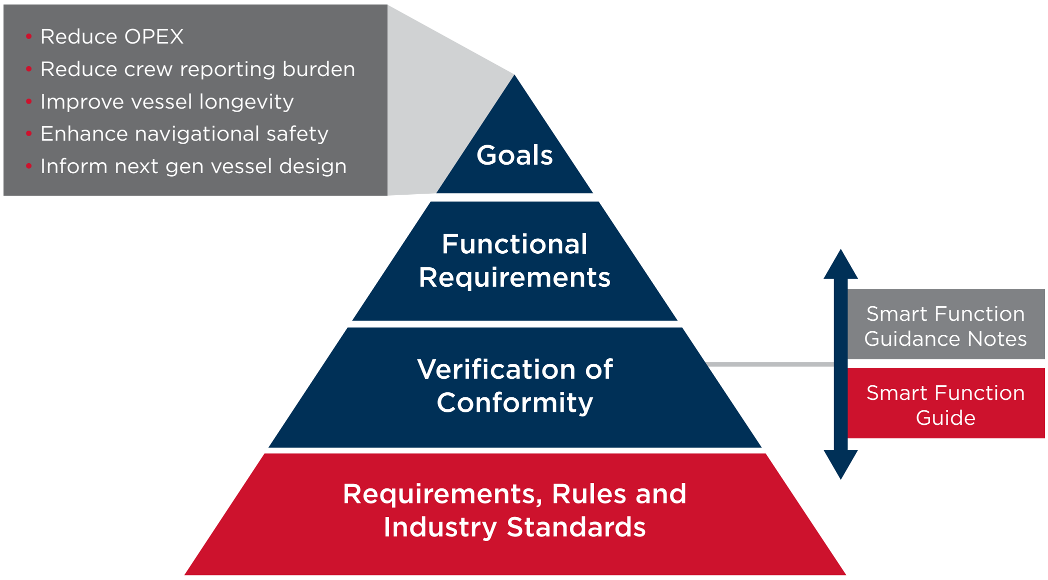


Figure 8: Goal-based Smart Function Framework

IIn addition to the Smart Guidance Notes, ABS has published the *Guide for Smart Functions for Marine Vessels and Offshore Assets* (the Smart Guide) to define the risk-informed technical requirements for Smart Function system(s) and offer optional SMART class notations and Class Record recognition (refer to Figure 8 for the coverage of the Guidance Notes and the Guide).

As data-centric applications, the Smart functions, as defined in the Smart Guidance Notes and the Smart Guide, rely on quality data and data analytics to derive health and performance conditions and assist decision making. Data quality assessment, monitoring and control is a key component for the Smart function implementation. The Verification and Validation (V&V) requirements for data quality, as summarized in the Smart Guide, are to be satisfied for the optional SMART notations and the relevant services so that ABS can trust the health, performance, or situational awareness related assessments coming from the use of the data and the employed Smart functions and services for class decision making. ABS, as a classification society, will independently verify and validate the system and procedure implemented for data quality assessment, monitoring and control by owners, operators, vendors, as well as third parties.

7 SUMMARY

With the rapid rise of digital technologies, connectivity and data processing and analytics, data-centric applications, such as Smart technology, are becoming increasingly common in the marine and offshore industries. To ensure a reliable and efficient data flow and accurate analytics outcomes, data quality must be managed throughout the data lifecycle.

Data quality assessments, monitoring and control are the key elements in the data quality management cycle. The main purpose is to ensure high quality data that fits the needs of data analytics, enabling data-driven functions and decision making.

This advisory provides information on the recommended data quality assessment tiers in alignment with the data quality categories as specified in ISO 8000-8:2015. To assess the data quality in each tier, two common approaches (i.e. top-down and bottom-up), along with the detailed procedures and activities of data quality assessments and controls are defined. The main capabilities of the data quality tools on the market and relevant techniques are reviewed.

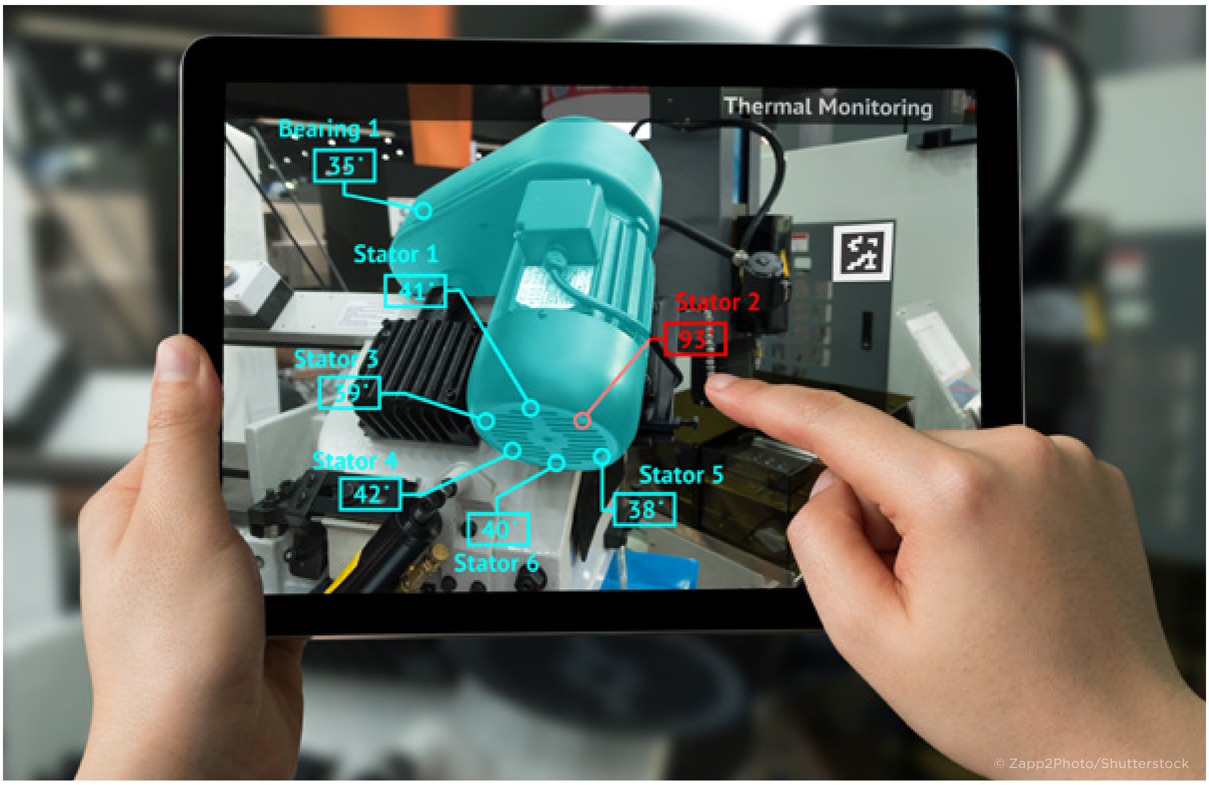
To get value out of the data, there are four important aspects which need to be managed properly within organizations:

i) Define data quality expectations: Data quality validation rules need to be defined clearly and documented consistently based on application goals and requirements, against which data quality can be validated. Data quality validation rules need to be reviewed continuously. The review process can help determine if the rules need to be refined.

ii) Develop a measurement system: A hierarchy of data quality measurement methods is recommended for measuring data conformance to the defined rules. Data quality results should be quantified and reported at multiple hierarchical levels (i.e., data quality metrics, dimension and overall levels)

iii) Identify improvement opportunities: The root causes of data issues need to be investigated to determine why and where the data defect originated. A remediation and improvement plan should be developed to address the root causes and to prevent the issue from recurring.

iv) Establish controls and report conformance to requirements: The operational data quality controls help prevent root causes from recurring and eliminate simple errors from occurring. An operational data quality control plan should be defined upfront. Data quality reporting helps the data consumers understand the ongoing condition of the data against the defined rules.



APPENDIX 1:
COMMON DATA QUALITY DIMENSIONS FOR MARINE IOT DATA

The data quality dimensions are measurable characteristics of data. They provide a vocabulary for defining data quality requirements and can be used to define results of initial data quality assessments as well as ongoing measurement. Regardless of the names that are used, dimensions focus on whether there is enough data (completeness), whether the data is right (accuracy, validity), how well the data fits together (consistency, integrity, uniqueness), whether the data is up to date (timeliness), accessible, usable and secure. Table 2 contains definitions of a set of common data quality dimensions mapped to their measuring methods, as well as the data quality categories as defined in ISO 8000-8. This table can be used as a go-by in the creation of an initial set of metrics to be used by an organization in measuring data quality dimensions.

Table 2: Common Dimensions used for Data Quality Assessment

Dimension	Description	Measurement	Unit of Measure	Target Questions	ISO 8000-8 Data Quality Category
Consistency	The extent to which data content and format are consistently represented within a data set and between data sets	i) Analysis of pattern consistency by using the presented structure and format within a column. – Structural Consistency ii) Analysis of the sampling rates in the same time series dataset. – Semantic Consistency	Percentage of consistent data (or inconsistent data)	Is data consistent within and across data sets?	Syntactic Semantic
Validity	The degree to which data conforms to the syntax (format, type, range) of its definition	i) Comparison between the data and the metadata or documentation for the data item, such as the allowable types (string, integer, floating point etc.), the format (length, number of digits, etc.). – Structural Validity ii) Validating the data value by comparing it to the defined domain of value (minimum, maximum or contained within a set of allowable values). – Semantic Validity	Percentage of data items deemed valid (or invalid)	Is the data consistent with its definition and requirements stated by the metadata?	Syntactic Semantic
Precision	The degree to which data shows the same level of details of the data value. Numeric data may need accuracy to several significant digits. For example, rounding and truncating may introduce errors where exact precision is necessary.	Analysis of the level of data accuracy, and rounded measurement values.	Percentage of data with accepted precision (or unaccepted precision)	Does the data have the necessary level of detail?	Syntactic
Completeness	The proportion of stored data against the potential of "100% complete".	i) A measure of the absence of blank (null) values or the presence of non-blank values. ii) Analysis of the conditions under which the missing value may/may not be valid for a certain data element.	Percentage of records with valid data (or invalid data)	Is all necessary data present?	Semantic

Dimension	Description	Measurement	Unit of Measure	Target Questions	ISO 8000-8 Data Quality Category
Uniqueness/ Deduplication	The degree to which data has no duplicate records within a data set. Nothing will be recorded more than once within the dataset based on how that data element is identified. For time series data, a data record with same values but different time stamps are not considered a duplication.	Analysis of the number of the given data element as assessed in the "real-world" compared to the number of records of the data element in the dataset. The real-world number of the data element could be either determined from a different and perhaps more reliable data set or a relevant external comparator.	Percentage of unique data (or duplicated data)	How many unique values are found for a given data element across all records? Are there duplicates? Should there be?	Semantic
Accuracy	The degree to which data correctly describes the physical parameter or event.	Accuracy is difficult to measure since the true physical value is typically unknown. Most measures of accuracy rely on comparison to a data source that has been verified as accurate.	Percentage of data entries that pass the data accuracy rules (or fail the data rules)	Does data reflect the real-world objects or a verifiable source?	Semantic
Integrity (Coherence)	The degree to which the intended relationship exists between the data in one column and the data in another column of the same or different data sets	Assessment of the intended relationship to ensure that the data in one column of a table can be traced and connected to data in another column of the same or different table.	Percentage of data missing important relationship	Is there any data missing important relationship linkages?	Semantic
Plausibility	The extent to which the compensated values (e.g. interpolated and corrected values) are used with the analytics algorithm. The methods of correction and rationale need to be recorded. An audit trial needs to be provided for monitoring the history of changes.	Assessment of the percentage of the compensated values used for data analytics.	Percentage of compensated values	Is there any compensated value used instead of real measurements?	Semantic
Timeliness	The degree to which data represents reality from the required point in time.	Comparison of the time between when information is expected (standard timestamps) and when it is readily available for use (actual timestamps). Time difference	Percentage of time difference	Is the data available at the time needed?	Pragmatic
Currency	The degree to which data is up to date. Data currency measures how "fresh" the data is, as well as correctness in the face of possible time-related changes.	Analysis of the conditions at which the data elements are expected to be stuck for a defined time period, and the expected frequency at which different data elements refreshed.	Percentage of stuck value	Are the data values the most up-to-date version of the information?	Pragmatic
Accessibility	The extent to which information is available, or easily and quickly retrievable.	Assessment of how easy it is to acquire data when needed, how long it is retained, how access is controlled.	User surveys through questionnaires or interviews. (e.g. If a SQL database is provided by the 3rd-party data supplier)	Is the data easily and quickly retrievable?	Pragmatic

APPENDIX 2:
COMMON DATA QUALITY RULES, METRICS AND DIMENSIONS FOR DATA QUALITY VERIFICATION AND VALIDATION

Data quality rules provide the foundation for operational management of data quality. Tables 3 through 5 give samples of the common data quality rules, measurable metrics and dimensions in accordance with the specific data quality problems. Each table formulates a structured way to verify and validate data quality in alignment with the three categories of data quality described in ISO 8000-8:2015.

Table 3: A Sample of Syntactic Data Quality Rules and Dimensions

Data Quality Issue	Problem Description	Syntactic Rule	Measurement (Data Quality Metrics)	Data Quality Dimension
Invalid Data Type	Invalid data type for the same data element, e.g. a text string found in a list of floating numbers.	Data type for the same data element must conform to metadata of its definition (e.g. text, number or date/time).	Percentage of invalid/inconsistent data values	Structural Validity
Different Data Format/Pattern	Different data format/pattern for the same data element, e.g. different format in a timestamp that makes data manipulation/comparation difficult.	Data format/pattern for the same data element must conform to metadata of its definition (e.g. DD.DD, DDDD-DD-DD DD:DD, where D represents a digit).		Structural Consistency
Values out of Range (unreasonable data values)	Data values are out of range for the domain under observation, e.g. value spikes or sudden changes which are implausible (99999, -99999) for the domain.	Data values for the same data element must conform to metadata of its definition (e.g. the assigned value is within a defined numeric, lexicographic, or time range).		Validity
Different Data Accuracy	Different level of data accuracy for the same data element (i.e. significant digits: number of digits on the right of the decimal point).	The level of data accuracy for the same data element must be same.		Precision

Note: The list of syntactic quality rules specified above does not exclude use of additional or alternative sets of syntactic quality rules.

Table 4: A Sample of Semantic Data Quality Rules and Dimensions

Data Quality Issue	Problem Description	Semantic Rule	Measurement (Data Quality Metrics)	Data Quality Dimension
Missing Data Values	There are gaps in the time series data.	The missing values under a specific condition are unacceptable.	Percentage of errors in data / population sample	Completeness
Duplicated Data Records	The data element is recorded with the same values more than once within the dataset that is unacceptable to what it represents in the real-world, e.g. duplicated timestamps.	The data element must have a unique representation in the real-world.		Uniqueness
Inaccurate Measurement	The value is slightly wrong which might result in the detection of a wrong trend etc. e.g. signal noise.	The measured data value shall match the property value for the real-world object it represents.		Accuracy
Diverging Sampling Rate	Different sampling rates in the same time series (same data source) can lead to problems (e.g. irregular timestamps).	The sampling rates in the same time series (same data source) need to be consistent.		Semantic Consistency
Divergent Despite High Correlation	Values which are normally correlated behave unexpectedly.	The intended data relationship linkage must be retained within the dataset.		Integrity
Forced / Calculated Value	Compensated values are used instead of real measurements.	Minimize the corrected (e.g., interpolation) data values which may reflect the assumptions made and no longer represent reality.		Plausibility
Wrong Timestamps or Wrong Timestamp Order	Timestamps are mismatched to the expected time (e.g., there are gaps/redundancies in timestamps); Timestamps are not in chronological order.	The recorded time shall match to the expected (standard) time.	Percentage of mismatched timestamps	Timeliness
Data not updated (stuck values)	Data is not up to date. Sensor might still display old values (e.g. defective sensors for which the data value is out of calibration).	Data must be updated with time.	Percentage of stuck values	Currency

Note: The list of semantic quality rules specified above does not exclude use of additional or alternative sets of semantic quality rules.

A Sample of a Pragmatic Data Quality Rule and Dimension

Data Quality Issue	Problem Description	Pragmatic Rule	Measurement (Data Quality Metrics)	Data Quality Dimension
Missing Foreign Keys	Foreign Keys are missing, which are used to identify the referential relation between tables in a SQL database.	The data is easy and quick to retrieve from a database (e.g., Foreign keys need to maintain referential integrity in SQL database).	User surveys (e.g. questionnaires)	Accessibility

Note: The pragmatic quality dimension specified above does not exclude use of additional or alternative sets of dimensions.

APPENDIX 3:
DEFINITIONS

Data: reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing [Source ISO/IEC 23822015].

Data Element: Unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes [Source ISO/IEC 11179-31994].

Data Error: Non-fulfilment of a data requirement [Source ISO 8000-2, 2017].

Data Quality: Degree to which a set of inherent characteristics of data fulfils requirements [Source ISO 8000-2, 2017].

Data Quality Dimension: A data quality dimension that represents the measurable feature or characteristic of data. Each dimension consists of a single or a set of data quality metrics.

Data Quality Management: coordinated activities to direct and control an organization with regards to data quality [Source ISO 8000-2, 2017].

Data Quality Metric: A data quality metric is an indicator that represents the data quality performance of data records (sensors) in a data set.

Data Quality Validation Rule: A data quality rule is defined to describe data quality requirements for determining/validating conformance of data to expectations.

Data Set: Logically meaningful grouping of data [Source ISO 8000-2, 2017].

Internet of Things (IoT): Infrastructure of interconnected entities, people, systems and information resources together with services which processes and reacts to information from the physical world and virtual world [Source ISO/IEC 209242018].

Metadata: Data that defines and describes other data [Source ISO/IEC 11179-12004].

Pragmatic Data Quality: The degree to which data is appropriate and useful for a particular purpose [Source ISO 8000-8, 2015].

Requirement: Need or expectation that is stated, generally implied or obligatory [Source ISO 8000-8, 2015].

Semantic Data Quality: The degree to which data corresponds to what it represents [Source ISO 8000-8, 2015].

Syntactic Data Quality: The degree to which data conforms to the specified syntax, i.e., requirements stated by the metadata [Source ISO 8000-8, 2015].

Time-series Data: Sequence of data values which are ordered in time.

Transaction Data: Data that represents the completion of a business action or a course of action [Source ISO 8000-2, 2017].

Validation: Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled [Source ISO 90002015].

Verification: Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled [Source ISO 90002015].

APPENDIX 4:
REFERENCES

- DAMA International. (2010). *The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK) – First Edition*. USA: Technics Publications, LLC.
- DAMA International. (2017). T*he DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK) – Second Edition*. USA: Technics Publications, LLC.
- ASQ. (2019). What is the Plan-Do-Check-Act (PDCA) cycle? Retrieved from American Society for Quality: https://asq.org/quality-resources/pdca-cycle
- ISO 8000-8:2015. (November 2015). *Data quality – Part 8: Information and data quality: Concepts and measuring.*
- ISO/TS 8000-1:2011. (2011). Data Quality – Part 1: Overview.
- Zhang, RJ; Jayawardene, V; Indulska, M; Sadiq, S; Zhou, XF. (2014). A Data Driven Approach for Discovering Data Quality Requirements. ICIS 2014: 35th *International Conference on Information Systems.* Auckland, New Zealand: University of Auckland Business School.
- ABS *Guidance Notes on Smart Function Implementation*
- ABS *Guide for Smart Functions for Marine Vessels and Offshore Units*

CONTACT INFORMATION

NORTH AMERICA REGION
1701 City Plaza Dr.
Spring, Texas 77389, USA
Tel: +1-281-877-6000
Email: ABS-Amer@eagle.org

SOUTH AMERICA REGION
Rua Acre, nº 15 - 11º floor, Centro
Rio de Janeiro 20081-000, Brazil
Tel: +55 21 2276-3535
Email: ABSRio@eagle.org

EUROPE AND AFRICA REGION
111 Old Broad Street
London EC2N 1AP, UK
Tel: +44-20-7247-3255
Email: ABS-Eur@eagle.org

MIDDLE EAST REGION
Al Joud Center, 1st floor, Suite # 111
Sheikh Zayed Road
P.O. Box 24860, Dubai, UAE
Tel: +971 4 330 6000
Email: ABSDubai@eagle.org

GREATER CHINA REGION
5th Floor, Silver Tower
No. 85 Taoyuan Road, Huangpu District
Shanghai 200021, P.R. China
Tel: +86 21 23270888
Email: ABSGreaterChina@eagle.org

NORTH PACIFIC REGION
11th Floor, Kyobo Life Insurance Bldg.
7, Chungjang-daero, Jung-Gu
Busan 48939, Korea, Republic of
Tel: +82 51 460 4197
Email: ABSNorthPacific@eagle.org

SOUTH PACIFIC REGION
438 Alexandra Road
#08-00 Alexandra Point, Singapore 119958
Tel: +65 6276 8700
Email: ABS-Pac@eagle.org

TX 01/20 19325