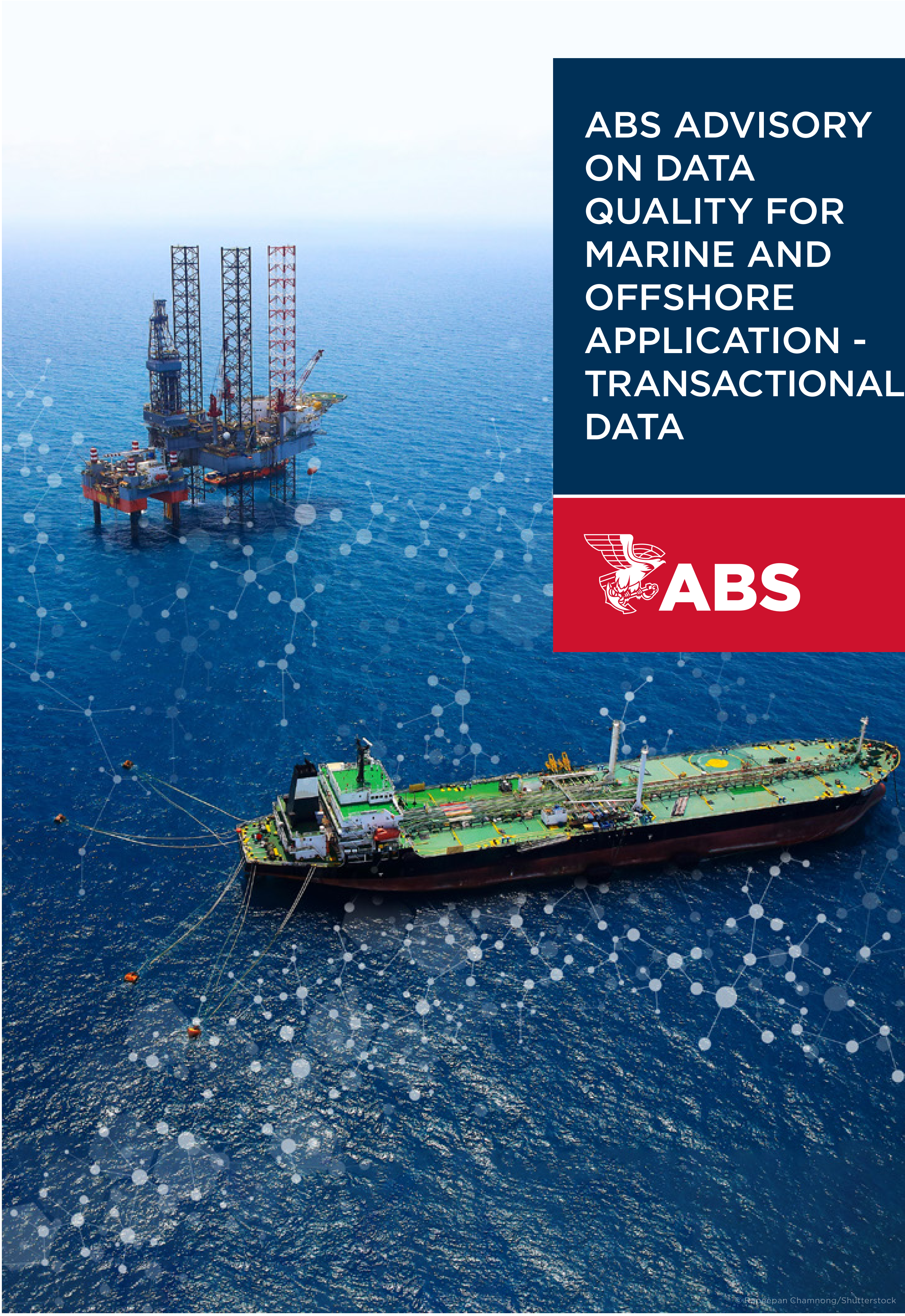
ABS ADVISORY ON DATA QUALITY FOR MARINE AND OFFSHORE APPLICATION - TRANSACTIONAL DATA

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

Data-driven models are increasingly being used in marine and offshore applications to monitor and predict asset conditions and provide decision support. Reliable data analytics and better decisions are built upon trusted data. To ensure data integrity and availability, the quality of data must be planned, assessed, monitored and controlled as part of the data life cycle in marine and offshore digital applications (as shown in Figure 1).

With the advent of data acquisition techniques and data-driven applications, the size of data collected for marine and offshore applications has grown drastically. The collected data has complex data structures and various data-type characteristics. Transactional data and time-series data are the two primary data types commonly generated during marine and offshore operations from onboard sensors, automation systems, crew input, inspection and survey. The marine and offshore industry should leverage both transactional and time-series data for generating asset condition and performance insights and supporting operational and asset integrity management decisions.

This document focuses on transactional data and aims to provide the industrial best practices on transactional data quality assessment, assurance and control. This advisory, together with the ABS *Advisory on Data Quality for Marine and Offshore Application – Time Series Data* (ABS, 2019) that addresses time-series data quality, provides the industry a comprehensive framework and guidance on data quality assessment, monitoring and control in marine and offshore applications.

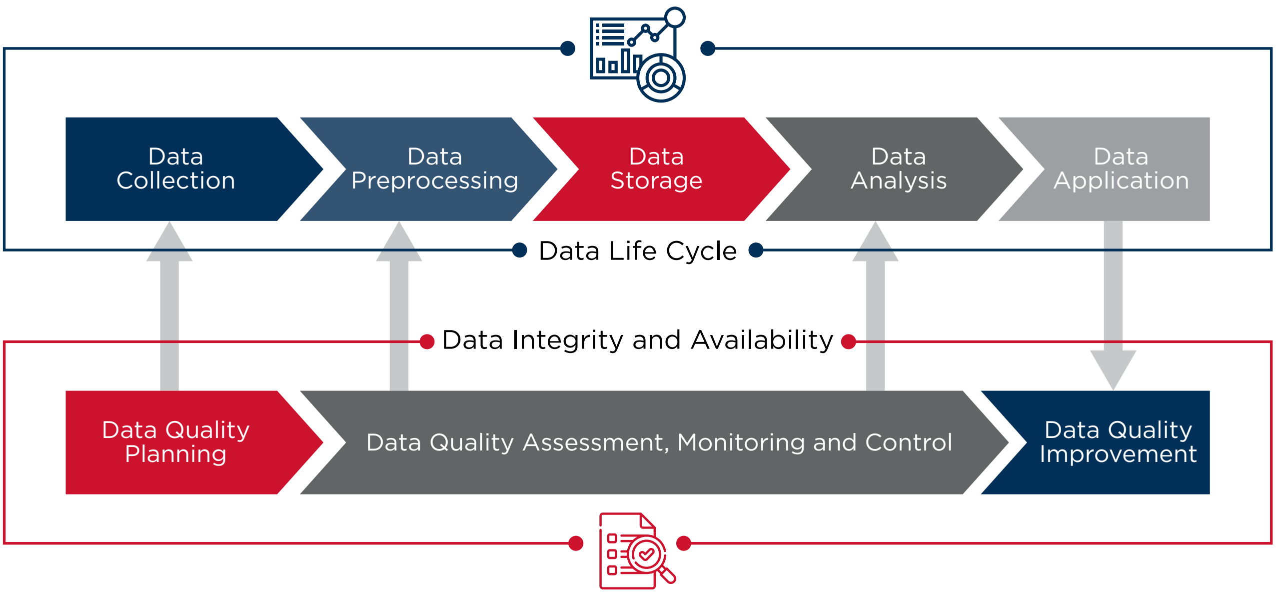


Figure 1: Data Quality Assurance throughout Data Life Cycle

1.2 DEFINITION OF TRANSACTIONAL DATA

In this advisory, the transactional data refers to the type of discrete, time-stamped business and device data generated by maritime operations, condition and performance monitoring, inspection, repair and maintenance activities (e.g., maintenance data, spare parts data, survey reports, noon reports and event data captured by sensors depending on the defined event-triggering threshold).

Compared to time-series data, which refers to the time-stamped data collected over a certain period at a particular frequency, transactional data comes in more varied forms and less prescriptive structures (e.g., MS Office documents, PDF files, emails, images, audio and videos) as shown in Figure 2. The variety of forms and lack of data structures make it difficult to directly apply the predefined data quality assessment measurement metrics and dimensions and traditional data quality toolsets as specified in the ABS Advisory on Data Quality for Marine and Offshore Application – Time Series Data. Although many applications implement ad-hoc data quality assessment and control for transactional data because of its characteristics, a systematic and consistent approach is beneficial and can better aid the data quality.

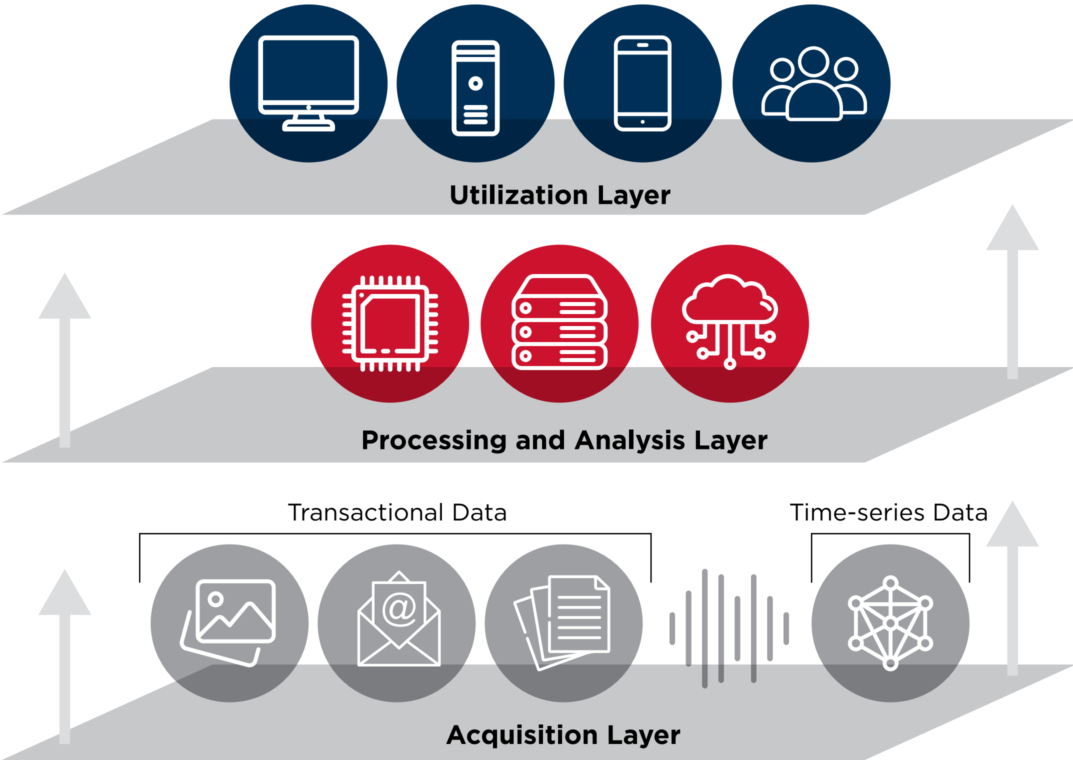


Figure 2: A Typical Data Management Architecture – IoT Data Flow

1.3 GOAL AND SCOPE

The main objective of this Advisory is to propose a systematic and consistent data quality assurance and control methodology to secure and improve transactional data quality throughout the data management architecture (see Figure 2). It is understood that if a company can manage the data quality of each dataset at the time when it is received or created, the data quality is naturally guaranteed.

First, the international standards relevant to transactional data quality are introduced and briefly reviewed. Second, the typical categories of the transactional data, the common data quality issues, available quality control practices and the potential data quality impacts for marine and offshore applications are summarized. Finally, a framework and guidance on how to plan, assess, monitor, and improve the data quality are detailed based on the quality improvement cycle 'Plan-Do-Check-Act' as defined in ISO 8000-61 (ISO 8000-61:2016, 2016). To make this methodology more practical than theoretical, a series of key activities and expected output generated by each activity are identified.

Section 1 describes the background of why transactional data quality requires a special consideration in the marine and offshore industry. Section 2 introduces the industrial and international data quality standards relevant to transactional data. Section 3 identifies the typical categories of transactional data and available quality control practices. Section 4 explores the common data quality issues associated with each of the transactional data categories and their potential impact on marine and offshore applications. Section 5 presents a data quality management and control procedure, which considers the best data quality practices for improving the quality of transactional data aligned to ISO 8000-61:2016. Section 6 describes ABS' role in the verification and validation of data quality. Section 7 provides a summary of the transactional data quality advisory. Finally, the Appendices section summarizes common transactional data quality issues and the recommended actions.

2 ISO DATA QUALITY STANDARDS

2.1 ISO 8000-61:2016

This standard (ISO 8000-61:2016 Data quality – Part 61: Data quality management: process reference model) provides a process approach to data quality management in a holistic way, overcoming the difficulties of isolated data quality activities. The basic structure of the data quality management process is composed of Implementation, Data-related Support and Resource Provision as shown in Figure 3 (ISO 8000-61:2016, 2016). The central process is Implementation which represents a continuous quality improvement model consisting of four repetitive sub-steps based on the 'Plan – Do – Check – Act' (PDCA) cycle (ASQ, 2019). Furthermore, this standard also defines the lower-level processes for data quality management as shown in Figure 4 (ISO 8000-61:2016, 2016). Each element is described by a purpose, outcomes and activities that are to be applied for the assurance of data quality.

This process approach is applicable to manage the quality of digital data sets including both structured and less structured data. However, the scope of this standard simply provides a process reference model for managing data quality and cannot serve as a methodology for data quality management. The process approach and PDCA model are done at a very high level, which are more theoretical than practical and need to be supplemented by detailed methods or procedures tailored for marine and offshore applications to achieve the outcomes of the defined processes.

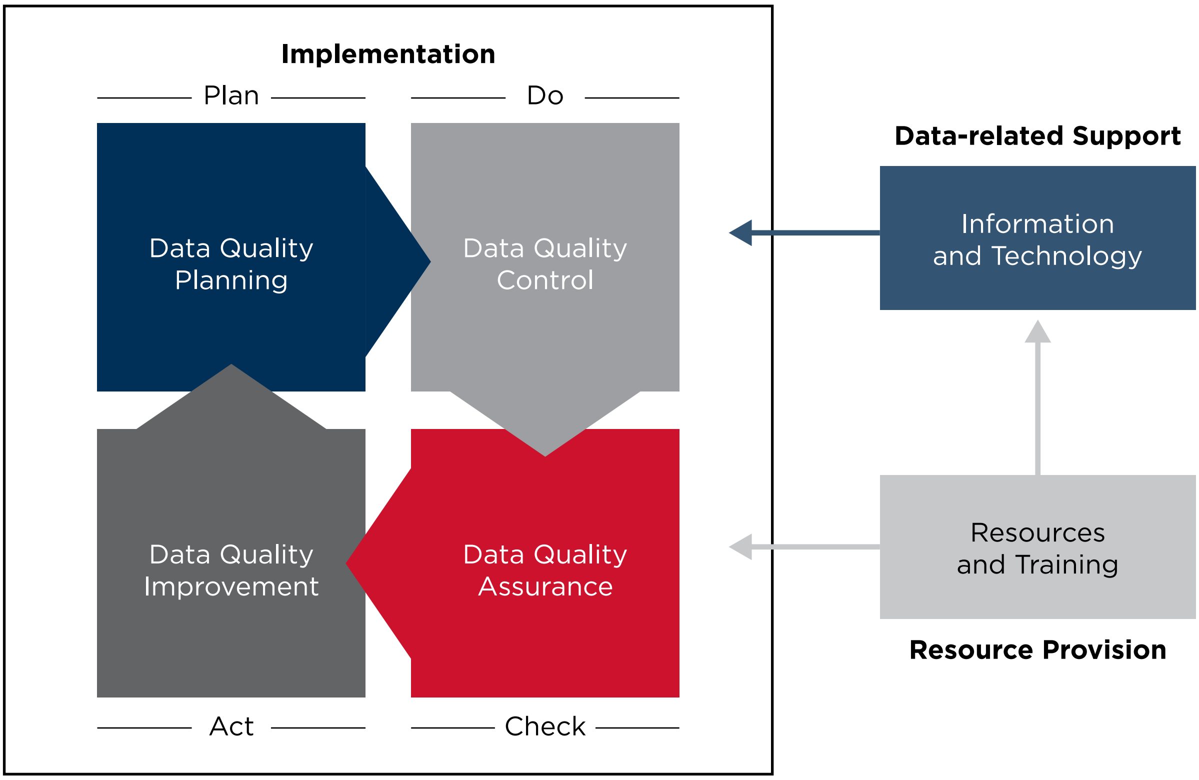


Figure 3: Basic Structure of Data Quality Management

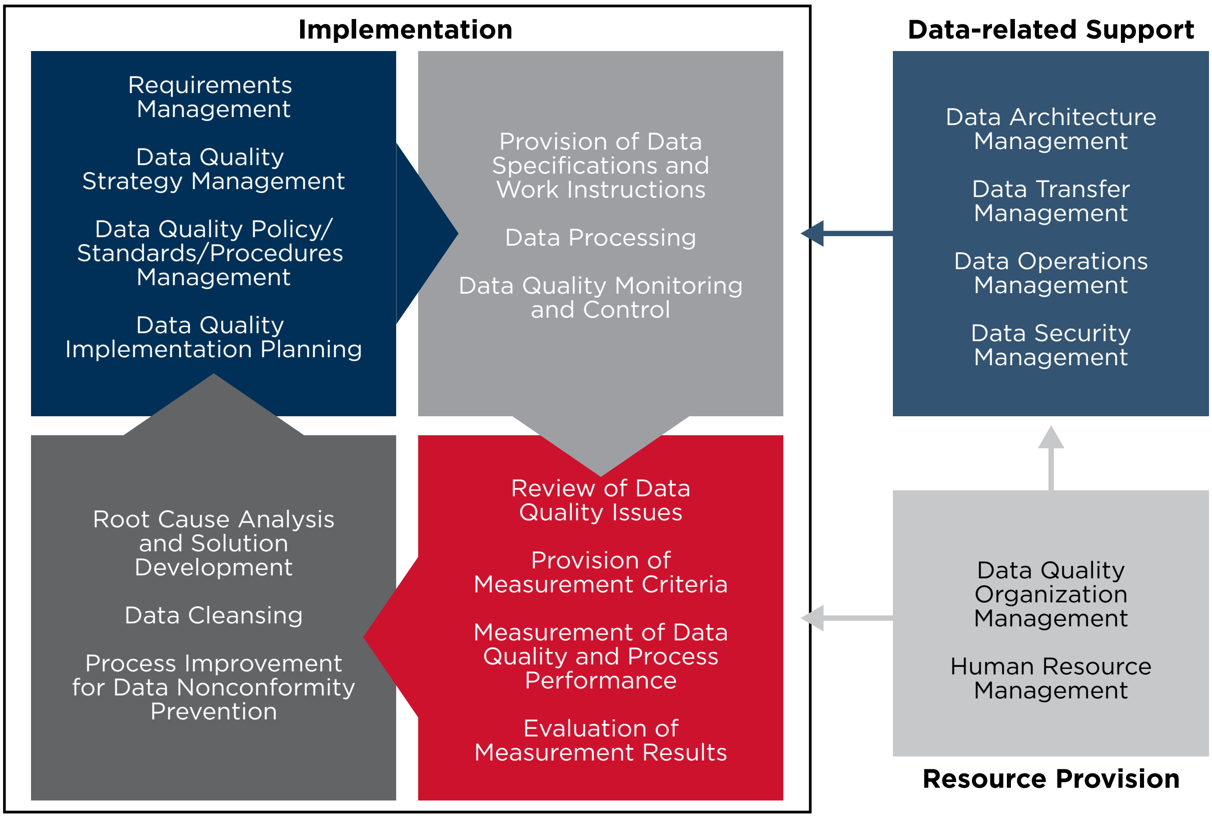


Figure 4: Detailed Structure of Data Quality Management

2.2 ISO 14224:2016

ISO 14224 (Petroleum, petrochemical and natural gas industries – Collection and exchange of reliability and maintenance data for equipment) provides a systematic approach for data collection and exchange, resulting in improved quality of data (ISO 14224:2016, 2016). It focuses on the data required in various analyses and provides a standardized data collection format for facilitating the exchange of reliability and maintenance data throughout the operational life cycle. It also sets the foundation for a consistent tracking of failures and maintenance records, allowing the prioritization and implementation of corrective actions.

This standard describes data quality control and assurance practices and provides guidance for the user regarding the quality of reliability and maintenance data. Firstly, prior to the data collection process, a set of planning measures need to be completed to ensure that consistent and compatible data are obtained. Secondly, during and after the data collection process, standardization of data collection practices is defined. The key actions are highlighted below.

2.2.1 Reliability and Maintenance Data Collection Plan

1. Define the objective for data collection. Reliability and maintenance data can be applied in different areas and analyses; some examples are listed in Table 1.

SAFETY	LIFE CYCLE COST/ OPTIMIZATION/MAINTENANCE	GENERAL
• Quantitative risk analysis (QRA) • Risk-based inspection (RBI) • Safety integrity level (SIL) • Environmental- and social-impact assessment (ESIA)	• Life cycle cost (LCC) • Reliability centered maintenance (RCM) • Spare-parts analysis (SPA) • Failure mode, effect, and criticality analysis (FMEA) • Root cause analysis (RCA)	• Manning-resource planning (MRP) • Six sigma (6∑) • Fault-tree analysis (FTA) • Markov process analysis (MPA)

Table 1: Areas of Application and Types of Analyses (Annex D, ISO 14224:2016)

2. Investigate the source(s) of the data to ensure that adequate data quality is provided. The five main criteria used to measure data quality are as follows:
    - Completeness - fulfilling expectations and meeting the needs according to specification
    - Consistency - complying with standard definitions (reliability parameters, data types and formats)
    - Accuracy - accurate input, transfer, handling and storage (manually or electronic), provide a high statistical confidence level (sufficient population and adequate surveillance period)
    - Timeliness - time expectation for the accessibility of data
    - Relevancy - data are relevant to the intended purpose

3. Define the taxonomical information for each equipment unit. The systematic classification (taxonomy) enables reliability and maintenance data to be meaningful and comparable. A hierarchy of the taxonomy levels is shown in Figure 5 (Section 8.2, ISO 14224:2016). Levels 1 to 5 represent a high-level categorization that relates to industries and plant application. Levels 6 to 9 are related to the equipment unit (inventory) with subdivision in lower indenture levels corresponding to a parent-child relationship.

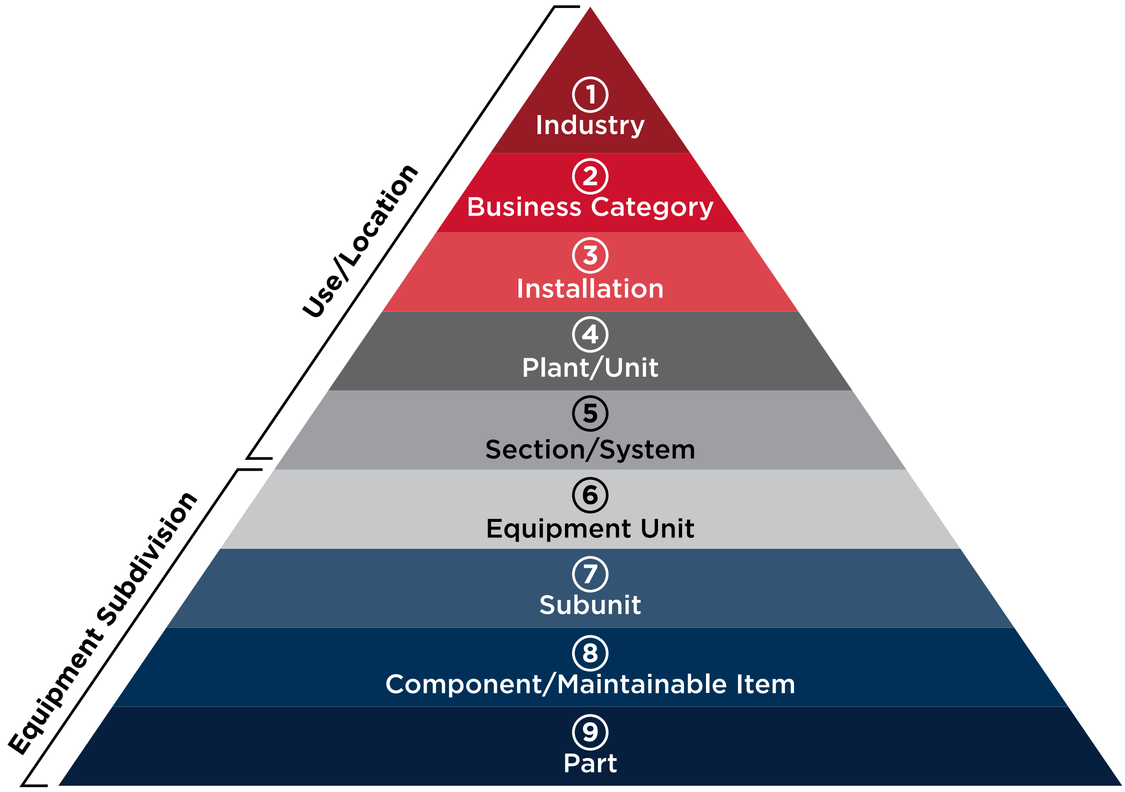


Figure 5: Taxonomy Classification with Taxonomic Levels

4. Identify the installation date, population and operating period(s) for the equipment considered. Timeline definition is presented in Table 2 (Section 8.3, ISO 14224:2016).

DOWNTIME	Planned downtime	Preventive maintenance	Preparation and/or delay
			Active preventive maintenance
		Other planned outages	Reserve
			Modification
	Unplanned downtime	Corrective maintenance	Undetected faults
			Preparation and/or delay
			Repair (item being worked on)
		Other unplanned outages	Shut-down, operational problems/ restrictions etc.
UPTIME	Operating time		Rundown
			Start-up
			Running
			Hot standby
	Non-operating time		Idle
			Cold standby

Table 2: Timeline Definitions

5. Define the boundaries for each equipment class, indicating what reliability and maintenance data is to be collected. A clear boundary description is necessary for collecting, merging and analyzing reliability and maintenance data from different industries, plants or sources (Annex A, ISO 14224:2016). This may be given by using a figure, a text definition or a combination of both. An example is shown in Figure 6.

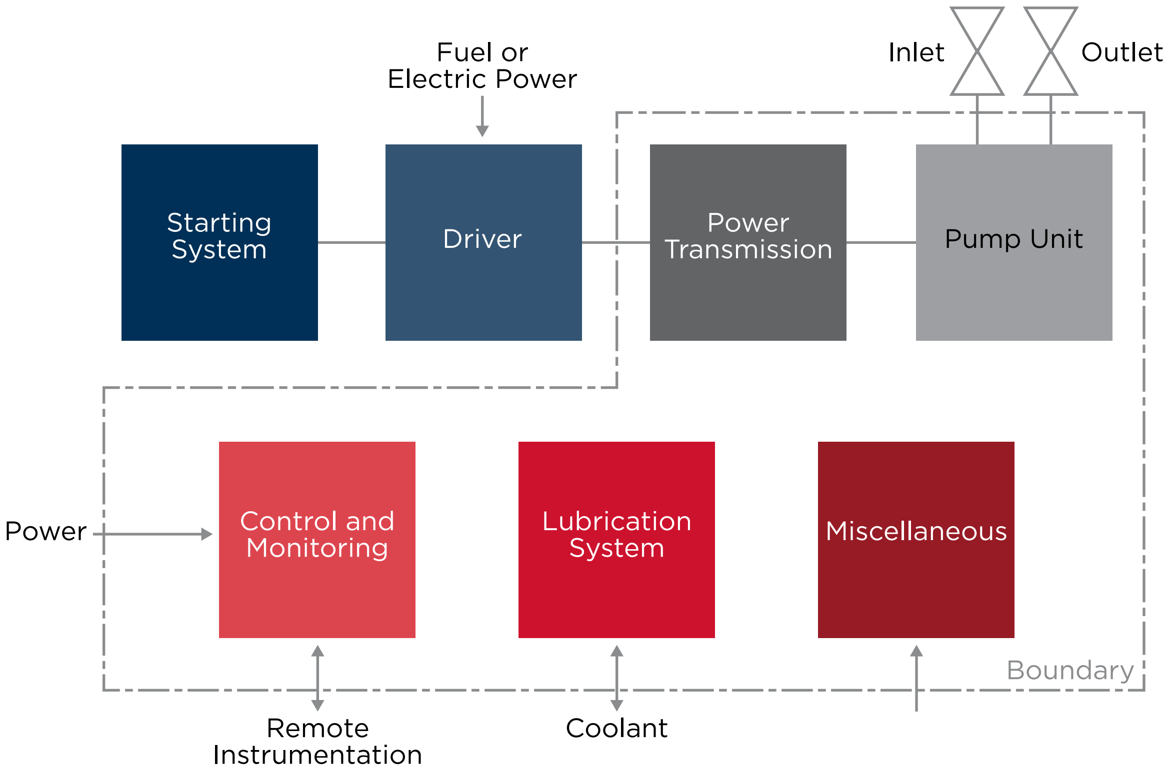


Figure 6: Example of Boundary Diagram (Pump)

6. Apply a uniform definition of failure and a method of classifying failures (Section 9.5, ISO 14224:2016). This standard establishes the minimum data needed. An example is given in Table 3.

IDENTIFICATION	FAILURE DATA	REMARKS
• Failure record • Equipment identification/ location	• Failure date • Failure mode • Failure impact on plant safety • Failure impact on plant operations • Failure impact on equipment function • Failure mechanism • Failure cause • Subunit failed • Component/maintainable item(s) failed • Detection method • Operating condition at failure • Operational phase at failure	• Additional information

Table 3: Failure Data to be Recorded

7. Apply a uniform definition of maintenance activities and a method of classifying maintenances (Section 9.6, ISO 14224:2016). There are two categories of maintenance defined as corrective and preventive. The data to be recorded are listed in Table 4.

IDENTIFICATION	MAINTENANCE DATA	MAINTENANCE RESOURCES	MAINTENANCE TIMES	REMARKS
• Maintenance record • Equipment identification/ location • Failure record	• Date of maintenance • Maintenance category • Maintenance priority • Interval (planned) • Maintenance activity • Maintenance impact on plant operations • Subunit maintained • Component/ maintainable item(s) maintained • Spare part location	• Maintenance man-hours, per discipline • Maintenance man-hours, total • Maintenance equipment resources	• Active maintenance time • Downtime • Maintenance delays/ problems	• Additional information

Table 4: Maintenance Data to be Recorded

8. Define the checks to be used in data quality verification. During and after the data collection, the data is assessed to verify consistency, reasonable distributions, proper codes and correct interpretations (Section 7.1, ISO 14224:2016). This standard identifies the minimum items to be verified.
    - The origin of data is documented and traceable
    - Data come from similar equipment type, technology and operating conditions
    - The equipment is relevant for the purpose (e.g., not outdated models)
    - The data comply with definitions and interpretation rules (e.g., definition of failure)
    - Recorded failures are within the defined equipment boundary and surveillance period
    - The information is consistent (e.g., consistency between failure modes and failure impact)
    - Data are registered in the correct format
    - Data volume collected gives acceptable statistical confidence
    - Operating and maintenance personnel are consulted to validate the data

9. Define a priority level for the completeness of data by a suitable method, such as using three classes of importance:
    - HIGH: compulsory data (coverage ≈100%)
    - MEDIUM: highly desirable data (coverage>85%)
    - LOW: desirable data (coverage>50%)

10. Define the level of detail of reliability and maintenance data reported and collected and link the level of detail closely to the production and safety importance of the equipment. Prioritize based on safety, production performance and/or other severity measures.

11. Prepare the plan for data collection, such as schedules, milestones, data collection sequence for installations and equipment units, surveillance periods to be covered, etc.

12. Plan data assembly and report and define a suitable method for transferring the data from source to reliability database. The typical data collection process consists of compiling data from different sources into one database where the type and the format of the data are predefined, as follows:
    - Address all available data sources and extract the relevant 'raw' data into intermediate storage
    - Interpret this information and translate it into the type and format desired for the target database (e.g., this is mostly done by manual interpretation)
    - Transfer the data from the source(s) to the reliability database using a suitable method (e.g., 'off-the-shelf' software has a robust conversion algorithm)
    - Carefully plan and test data collection methods before starting the main process, since these methods significantly impact the cost-benefit analysis for data collection

13. Train, motivate and organize the data collection personnel, ensuring in-depth understanding of the equipment, its operating conditions, ISO 14224 and the requirements for data quality.

14. Prepare a plan for quality assurance of the data collection process and its deliverables. ISO 14224 establishes, as a minimum, procedures for quality control of the data and for recording and correction of deviations (e.g., the verification of quality process during and after the data collection exercise).

15. ISO 14224 recommends performing a cost-benefit analysis of the data collection by running a pilot exercise before the main data collection phase is started.

16. Review the planning measures after a period of using the system.

2.2.2. Reliability and Maintenance Data Collection Process

Data collection is an investment. A cost-effective data collection process is recommended. Standardization of data collection processes is divided into three steps:

1. Initial data quality verification (before the main data collection): It is recommended to conduct a pilot exercise to verify feasibility of the actions established in the plans (as specified in Section 2.2.1). An initial data verification is required to identify any problems that require the planning measures to be immediately revised to avoid the collection of unacceptable data.

2. Data quality control (during the data collection process): A system for dealing with deviations encountered shall be established and problems solved as soon as possible (e.g., data correction and data cleansing).

3. Data improvement: All quality lessons learned during the planning and execution of the data collection effort are to be evaluated and summarized. Recommendations are then forwarded for improvement of the process.

2.3 SUMMARY

The marine and offshore industry seeks mature industry common practices for data quality assessment, assurance and control for execution. ISO 8000-61 provides a fundamental structure for the data quality management process, but it does not serve as a methodology for data quality management in a particular domain. ISO 14224 presents a comprehensive basis for the collection and exchange of reliability and maintenance data (i.e., a main source of marine transactional data as identified in Section 3). The quality of reliability and maintenance data is highly dependent on the method the data is collected and reported in the first place. This standard provides a systematic approach for standardization of data collection exercises, resulting in improved data quality.

In this respect, this Advisory will follow the process approach presented in ISO 8000-61 and the data standardization for reliability and maintenance data collection and exchange summarized in ISO 14224 to establish data quality management and control methodology and best practices for improving marine and offshore transactional data quality.

3 TYPICAL CATEGORIES OF TRANSACTIONAL DATA IN THE MARINE AND OFFSHORE INDUSTRY

Three typical categories of marine and offshore transactional data are identified for common data-driven applications, including reliability and maintenance data, manually input data and transactional sensor data. Each category can be further divided into sub-types, and the available data quality analyses suitable for each sub-type are also discussed.

3.1 RELIABILITY AND MAINTENANCE DATA

3.1.1 Computerized Maintenance Management System (CMMS) Data

It is common practice for equipment reliability and maintenance data to be stored in a system referred to as a Computerized Maintenance Management System (or CMMS software). This is widely used in the marine and offshore industry to store equipment lists, their maintenance task instructions, task schedules, track maintenance tasks, such as work order tracking, budget tracking, automatically generated preventive maintenance schedules, advanced reporting and asset management.

The main purpose of CMMS is to serve the vessel's operators as the 'corporate memory' to provide replacement crews a readily available record of the equipment onboard, identify past problems, generate work orders, inform the maintenance crew and track assigned and completed tasks. CMMS data contains complete records regarding the equipment's failure history, the planned maintenance work and the maintenance actions taken.

A CMMS is able to provide readily accessible information for use in optimizing the timing of equipment overhauls and inspections by assessing the cost of machinery breakdown repairs versus preventative maintenance, reducing equipment failures versus the life cycle costing of spares and upgrading maintenance programs for all types of operating facilities. Technicians provide preventive maintenance with continuous operation instead of reacting to unplanned failures and downtime.

CMMS data may also be used to verify regulatory compliance. Class needs to verify the equipment maintenance records for vessels enrolled in the Preventative Maintenance Program (PMP).

CMMS stores comprehensive machinery condition records, and the CMMS data is also widely used to understand the typical failure behavior of equipment through a well-established RAMS (Reliability, Availability, Maintainability and Safety) analysis through a statistical analysis of equipment failure data to derive the failure probability or reliability distribution functions.

With the advent of big data and data analytics, data-driven and machine learning modeling becomes an important component and means for equipment condition and performance monitoring. Supplemental to equipment operational data, CMMS data provides records on equipment failure and maintenance that is critical for model training, testing and validation.

3.1.2 Spare Parts Inventory Data

The spare parts inventory data is a type of maintenance data which provides information for the current stock, cost, usage and location of spare parts on board or on shore. The demand for such systems arises from past equipment failure experiences. The purpose of spare parts inventory control is for maintenance. A CMMS can perform calculations to analyze and organize the spare parts data based on the user defined usage percentages. After the analysis, the inventory items can be categorized into different groups (e.g., high-, medium- and low-) in terms of stock volume and/or cost. Thereby, a plan to optimize inventory stocks for the most used parts can be created.

Typically, the cost of maintaining spare parts' inventories for capital-intensive industries such as maritime assets is significant. In addition, there are challenges in the procurement and ordering of the valuable marine spare parts, which includes high transportation and inventory cost, but also, long lead times and possible obsolete parts that are no longer in production. Reducing these inventories can be very costly due to the substantial downtime cost for maritime assets.


© tsyhun/Shutterstock

3.1.3 Available Data Quality Analysis and Control Practices

There are few published works considering the quality assessment related to semi-structured and unstructured reliability and maintenance data. Case studies addressing the maritime industry are lacking. For equipment reliability (survival) analysis, there are two types of approaches applied for data quality analysis. The first one is a criteria-based method (M. R. Hodkiewicz, N. Montgomery, 2014) which allows users to evaluate the quality level of data required for the key asset-related reliability and maintenance decisions. This is a qualitative method, which is based on a set of descriptive criteria for each data element as predefined. For example, failure mode is one of the required data elements for a maintenance decision. There are four (4) different levels of data fitness for purposes associated with the data quality scores (ranging from 1 to 5) to assist in the assessment of the failure mode data quality. The maximum score obtained by most organizations for failure mode data is 3 by inferring failure modes from freeform text. The user can assess their data element against each of the criteria and produce a score. The score is then compared with the minimum required level for a specific decision.

The second approach is to outline the design philosophy of a data quality dashboard for reliability data, which is based on a set of key metrics (key performance indicators (KPIs), e.g., sampling size, observed time window, data field completeness, inconsistent names) suitable for reliability data quality analysis (Ralf Gitzel, Simone Turring, Sylvia Maczey, 2015) (Ralf Gitzel, Subanatarajan Subbiah, Christopher Ganz, 2018). The typical data quality problems are identified through case studies, and then a set of measurement formulas are developed to calculate the metrics. A prototype of the traffic light system is presented to show an immediate feedback regarding the suitability of the data as input for reliability analysis.

Reliability and maintenance data have complex data formats such as freeform text, and the total number of records is a variable within a certain period depending on the different operation and asset conditions. Compared to time-series data, it is difficult to apply the metrics-based analysis methods to quantify all the transactional data quality problems. For reliability and maintenance data, quality control for data collection and data entry is critical. The direct control approach can be applied to reduce data inaccuracies from the data sources and manual inputs, which may include:

- Identifying the source of data inaccuracies
- Standardizing manual input and avoiding freeform text
- Applying data validation for manual input
- Implementing crew assistance software and functions to reduce human errors
- Providing a data quality dashboard to continuously monitor the data quality

In addition to direct data quality control, an indirect approach to process and quality assurance is also important to improve the quality of data collection and data entry. These indirect approaches may include:

- Improving the organizational data quality maturity
- Promoting data quality awareness
- Providing adequate training
- Establishing quality assurance and control for data collection and data entry
- Ensuring the work is double-checked
- Providing a more crew-friendly working environment
- Implementing crew assistance software and functions to reduce crew workload

3.2 MANUALLY INPUT DATA

3.2.1 Noon Report Data

The ship noon report typically contains a comprehensive set of variables including ship position, draft and trim, average ship speed, daily fuel consumption, sailing direction, observed wind, wave and weather conditions and other relevant data recorded on a daily basis and manually entered by the ship's crew (e.g., captain or chief engineers). The reports are often required by charterers to monitor ship voyage performance and to confirm schedules are being maintained. The ship noon report data is widely used by shipping companies to assess vessel performance or estimated arrival times to the next port. Analyses of the vessel's performance data can be used to identify degrading trends and applied for a comparison of relative performance among identical vessels and different vessels for fleet-wide comparisons. Results can then be applied to vessel operating protocols so as to improve fleet operations or minimize fuel expenses (e.g., evaluate the effect of weather, hull fouling, trim and loading condition to the overall speed/powering performance of the ship). Primary measurement parameters (speed, power from either shaft torque and engine revolutions or fuel consumption) would be averaged over a certain period.

The noon report data quality depends on the skills of the people who prepared it and the subsequent calculations conducted on board. For example, the ship's crew may report the engine power through basic calculations using shaft rotational speed (RPM) and propeller torque and thrust estimates when no torque meters are installed. The calculation may not be accurate if the crew does not exactly follow the engine maker's recommended procedure, the propeller is fouled, deformed or has numerous pits. In addition, the noon report data may have some intrinsic problems: some of the data are instantaneous and collected directly from sensor readings (e.g., wind speed measurements by the anemometer), which may not reflect the average weather condition over the entire 24 hours. Often, daily reports are completed at noon local time. If a ship transits time zones during a voyage, the time period between a daily noon report may denote more or less than a 24- hour interval.

Noon reports provide a brief, time-averaged view on vessel operations when sensor-based continuous operation monitoring is not available. In addition to the voyage summary, noon reports may also be used for other purposes, such as fuel consumption monitoring. Vessel performance may be analyzed based on noon report data when other direct onboard measurements are not available. To improve noon report quality, it is important to identify the potential usage of the collected data with the recipients of the analyses and reports and possible data quality problems relevant to the noon reports.

Data collection methods based on continuous monitoring of ship parameters (auto-logging) are generally superior compared to use of traditional noon reports. Particularly when the frequency and density of data is high, applications such as trim-optimization, weather routing, fuel consumption and emissions monitoring are highly accurate. However, for cases where noon reports are used, good correlation to the auto-logging process may be possible especially if the time horizon is large enough.

3.2.2 Inspection and Survey Data

Inspections and surveys are carried out to verify the safety and operability features of the vessels are functional at the time of the surveys. Inspections of equipment systems and other assets are typically conducted by accredited service companies and trained personnel. For the periodical class surveys (annual, intermediate and special survey), surveyors need to conduct visual inspection, close inspection, NDT, and open the machinery/equipment (engine, rudder, tail shaft) to check the wear and tear, based on class rules and requirements. All survey findings are recorded in a specific database. Surveyors need to provide the survey reports and any mandatory or recommended actions to the client.

There is a variety of inspection and survey data types, such as gauging, vibration measurement, tank conditions, structural damages (crack, deformation, etc.), corrosion, etc. with various data formats (problem descriptions, images, measurement data, etc.). With the advent of innovative inspection technologies, more data types and huge data volumes become the norm for inspection and survey, such as videos from drones and remotely operated vehicles (ROVs) and 3D scanning data.

The activities of the preventive maintenance (either condition-based or planned maintenance) and the history of all maintenance tasks are recorded and stored in the CMMS by the operator or owner. Details of the CMMS data are to be reviewed by the surveyor during the implementation survey for verifying the regulatory compliance. Inspection and survey data/findings provide a comprehensive overview with many details regarding the asset's condition. In addition to compliance reporting, more and more condition assessment and prediction-relevant applications use inspection and survey data as the main data source. Understanding potential data quality issues to improve the quality of inspection and survey data is crucial for these applications.

3.2.3 Available Data Quality Analysis and Control Practices

The current research on noon report data quality focuses on the quality control, data enrichment and uncertainty measurement. The common method is the initial data quality screening and control. A set of validated data ranges have been predefined for data elements by the subject matter experts (SMEs) to identify and avoid data outliers. Data completeness is the most important quality metric needed to be measured for noon reports. In addition, various statistical and data mining methods, such as K-Mean, Self-Organizing Map, Outlier Score Base and Histogram Outlier Score Base, etc. are proposed in different literature to correct and enrich the noon report data quality (Ali Akbar Safaei, Hassan Ghassemi, Mahmoud Ghiasi, 2018). The underlying uncertainty in noon reports has been quantified by using a multiple linear regression model. The key finding is when compared with the continuous monitoring data (auto-logging) captured for the same vessel during the same operational period, the uncertainty level is increased (e.g., the confidence interval and standard error derived from the regression model) when noon report data is recorded manually, which reflects more scattered and inconsistent data and potential human errors.

The potential for variations due to human errors and subjective judgement exists when data collection processes are not automated. Although the use of various mathematical models (e.g., statistical, numerical and analytical models) is possible to correct erroneous noon report data, the fundamental principles for improving the quality of manually input data is to monitor and control the whole data collection and reporting process.

Similar data quality control principles apply to reliability and maintenance data for the noon reports and inspection and survey data. Some suggested strategies include:

- Adopting automated data collection and e-logging system
- Using pre-defined options and choices instead of free input
- Avoiding ambiguous definitions and terms
- Identifying source of data inaccuracies
- Applying data validation
- Using a data quality dashboard to continuously monitor the data quality

When manual input is unavoidable, data quality control, process and quality assurance in-direct approaches as described in Section 3.1.3 are also applicable.

3.3 EVENT DATA GENERATED BY SENSOR

In this Advisory, the event data refers to the events detected and captured by onboard sensor(s). Occurrence of the events may depend on the pre-defined event-triggering threshold or algorithm. For example, the slamming events detected through accelerometers and/or pressure transducers as described in the ABS Guide for Hull Condition Monitoring (ABS, 2020) are the events of interest. All the features of identified events should be stored in a computer-readable format.

3.3.1 Available Data Quality Analysis and Control Practices

Since the discrete events are detected through physical sensor data, the sensor data quality is critical and the relevant data quality analysis and control practices are appliable and can be referred to the existing framework and data quality rules summarized in the time-series data quality advisory (ABS, 2019). The most cost-effective way of optimizing data quality is through industry co-operation. Some solutions propose that owners, manufacturers and vendors should align their data quality management mechanisms with well-known international standards. For example, a methodology based on ISO 8000-61 for data quality management in sensor networks is introduced, which consists of four steps according to the 'Plan-Do-Check-Act' cycle (Ricardo Pérez-Castillo, Ana G. Carretero, Ismael Caballero, Moisés Rodríguez, Mario Piattini, Alejandro Maté, Sunho Kim, Dongwoo Lee, Sep. 2018). Moreover, some publications introduce a schema of data quality control throughout data management which is highly on demand. It consists of several aspects, such as data quality dimension selection, design of quality inspection, regulation of quality control standard and theories of the data usability, data autodetection and auto restoration (Dongmei Huang, Danfeng Zhao, Lifei Wei, Zhenhua Wang, Yanling Du, 2015). Furthermore, Internet of Things (IoTs) has been rapidly deployed in manufacturing, mining and construction industries. The quality of sensor data plays a vital role in IoT applications, as 'garbage-in-garbage-out' cannot render actionable decision-making. Some systematic reviews are presented which identify the different types of physical sensor data errors and the various data mining algorithms used for error detection and correction for sensor data. The current challenge is that it is difficult to compare these proposed methods and techniques for sensor data error detection and correction. An open-source benchmarking system is required for techniques that solve sensor data quality issues. The benchmark should provide datasets complete with all the different types of errors and a proper scoring system that uses the appropriate evaluation metrics to allow comparability of methods in terms of their performance to solve sensor data quality issues (Hui Yie Teh, Andreas W. Kempa-Liehr, Kevin I-Kai Wang, 2020).

4 DATA QUALITY ISSUES AND IMPACTS

Understanding the quality issues associated with transactional data are essential for assessing, monitoring and improving data quality. This section highlights the typical data quality issues and their potential impacts associated with the sub-types of transactional data as summarized in Section 3. The ABS *Advisory on Data Quality for Marine and Offshore Application – Time Series Data* lists the typical data quality metrics and dimensions suitable for time-series data, among which some are also applicable to transactional data. In addition, Appendix 1 to Appendix 3 illustrate the common data quality issues relevant to transactional data.

4.1.1 Issues Associated with CMMS Data and Impacts

The CMMS data can be used to understand the typical failure behavior of equipment through a well-established RAMS (Reliability, Availability, Maintainability and Safety) analysis. A core element of RAMS is the statistical analysis of equipment failure data to derive the failure probability or reliability distribution functions. Therefore, the data quality issues can mislead maintenance planning, failure forecasting, warranty planning and reliability optimization.

When the CMMS data is used to train and test the data-driven models for machinery condition and performance insights, in addition to the data quality itself, data unbalance is another common issue. Operation and maintenance data typically includes a large portion of normal operations while lacking certain operating modes and failure records. Special treatment of the data, such as down sampling on the normal operation data, is necessary before it can be applied for the model training. For a statistical analysis to be relevant, some equations have been proposed to estimate appropriate sampling size which depends on a standard error of the mean (SEM) and the mean time between failures (MTBF) (Ralf Gitzel, Simone Turring, Sylvia Maczey, 2015).

The typical data quality issues with CMMS data are summarized below, the details of which can be found in Appendix 1:

- Insufficient sample size of the failure data
- Lack of information on time to failure
- Inconsistency in data type or format
- Logical errors – failure events referring to an asset or functional location which does not exist. This is because the asset data does not contain a list of relevant failure events. A functional location is used to bridge both the failure event and an asset. The connection between asset and failure event is indirect. Over time, assets can be replaced, repaired, or overhauled and installed in other places
- Illogical order of dates
- Missing events – missing rows and missing columns
- Implausible reliability data is not wrong - Reliability data is artificially changed by service technicians or SMEs based on their domain judgements. For example, implausible dates (e.g., failure on the day of delivery), similar entries (e.g., two sites which had identical failures) and round figures (e.g., 5000 hours)
- Low in richness of failure information – For example, unstructured 'free-text' failure descriptions are used rather than a standardized failure code. The date of start of operation and the date of failure are not available, the 'related' dates (e.g., delivery date, or reporting date) are used instead of actual dates. These will affect the accuracy of the advanced reliability analysis

4.1.2 Issues Associated with Spare Parts Inventory Data and Impacts

Inaccurate spare parts inventory data can lead to poor spare parts management and maintenance planning and optimization, which may cause inefficient inventory storage and a shortage of parts when the material is required on board, leading to unplanned downtime and unforeseen penalty costs.

The typical data quality issues with spare parts inventory data are summarized below, and the details are also tabulated in Appendix 1:

- Manual data entry errors which cause an incorrect quality request for a material
- Improper equipment registers and lack of real-time tracking which missed the production information and schedule
- Lack of failure data to determine a spare part demand pattern during the use of the newly-issued equipment
- Wrong classification on equipment criticality (risk level) which underestimate the service level and downtime penalty costs for critical equipment


© moreimages/Shutterstock

4.1.3 Issues Associated with Manually input Data and Impacts

Experience shows that manually generated noon reports often contain large uncertainties resulting from daily averaging of widely varying weather conditions and increased probability of human error. Another typical human error may be caused by copying the previous report and making alterations. An erroneous noon report cannot provide much insight into the underlying causes of increased fuel consumption or overall deteriorating performance of a ship.

It requires crew training and continuous monitoring of the data collection process. The data quality is directly dependent on the quality culture and maturity level of the company.

Free texts are another common hurdle for manually generated reports for further data processing and analytics. It is hard to access and extract the data efficiently from the vast number and variety of manually generated reports. In this respect, reducing the free text input, creating standardized report templates (e.g., forcing the user to choose from a drop-down list) to limit the subjective input and building a suitable structure to partition and store the vast number and variety of reports are essential elements to improve the data query, increase the data usability and render decision support.

The typical data quality issues with manually input data are summarized below. Further details can be found in Appendix 2.

- The instantaneous data value is not reflected in the noon report
- Noon report data is recorded based on calculation rather than a direct measurement
- Subjective judgment/forced values are estimated by crew with subjective judgment, such as the wave conditions
- Inconsistent data format/pattern for the same data element exists
- Different levels of data accuracy for the same data element exists
- Missing events – missing rows and missing columns
- Data values are out of range for the domain under observation
- Timestamps are mismatched to the expected time (e.g., time gaps/redundancies exist) or are not in chronological order
- The text of the inspection and survey report is poorly constructed (spelling, grammar) which provides a completely different meaning from what the surveyor wants to express
- Low detail of survey information – the digital photographs are widely used for marine survey reports over the past 10 years. In some circumstances, surveyors only use captions adhered to the pictures instead of formal reports, lack of the supporting texts/sentences to explain the points that they want to make

4.1.4 Issues Associated with Event Data Generated by Sensors and Impacts

The quality issues of event data generated by sensors attribute to the typical sensor data errors including bias, missing values and inaccuracy (Hui Yie Teh, Andreas W. Kempa-Liehr, Kevin I-Kai Wang, 2020). An IoT application may have hundreds or thousands of sensors which produce vast amounts of data. The sensor data is a fundamental element of marine and offshore IoT architecture. The frequency of event data depends on the defined event-triggering threshold. Poor sensor data quality may lead to inaccurate situation and condition awareness. These quality issues must be detected, quantified, removed or corrected through IoT data flow.

Figure 2 shows an overview of the data flow of a typical IoT architecture, which consists of three layers, including: a data acquisition layer (i.e., data ingestion); processing and analysis layer; and utilization layer. The separation of the layers and complexity of IoT architecture can also cause potential data errors in each layer. For example, in the data acquisition layer, sensors must be tested and calibrated periodically. Sensors installed on board the marine assets are often subjected to extreme weather conditions, which may affect the readings and accuracy. In the processing layer for data transmission, the congested and unstable wireless communication links in sensor networks may cause data loss and corruption.

Industry-grade sensors are more accurate, stable and robust but they incur higher costs if the large and dense sensors networks are deployed for IoT applications. Today, many IoT applications in marine and offshore use low-cost sensors to reduce the initial investment although there may be a high penalty in data quality (e.g., data processing and correction). Since sensor data errors may be present and propagated in all layers, a general trend and the best practice of big data/IoT application is to shift the data processing and quality check to the data acquisition layer, so that only data with good quality is passed to the central system. Edge data processing and enrichment will be the real-time solutions for the processing of sensor data quality for IoT data management (Informatica, IoT Data Management: The Rise of Industrial IoT and Machine Learning, 2020) especially when 5G becomes widely available.

The data quality issues with the event data generated by sensors are summarized below which refer to the ABS Advisory on Data Quality for Marine and Offshore Application – Time Series Data. The details can be found in Appendix 3

- Missing data – The sensor stops providing any reading on its interface which results in absence of values/events. This may be caused by data disruption due to the constraints from data compression and security encryption losses
- Stuck/jammed data – The sensor reading gets jammed and stuck in some incorrect value
- Bias – The observed sensor reading deviates from the expected value by a constant offset or produces a temporal delay (e.g., sample selection bias, time-period bias)
- Invalid data type – Invalid data type for the same data element. This may be caused by data disruption due to the constraints from data compression and security encryption losses
- Different data format/pattern – Different data format/pattern for the same data element
- Wrong timestamps or illogical chronological order – Timestamps are mismatched to the expected timeline. Timestamps are not in chronological order

5 DATA QUALITY MANAGEMENT PROCEDURE FOR TRANSACTIONAL DATA

Based on the overview of IoT data flow as shown in Figure 2, it is necessary to assure and maintain data quality levels in every stage of the data life cycle.

To establish the best practices and provide recommendations to the industry on transactional data quality assurance and control, from Appendix 1 to Appendix 3, different types of transactional data quality issues are identified and mapped to the data quality improvement cycle 'Plan-Do-Check-Act' as defined in ISO 8000-61 (refer to Figure 3). It is intuitive that transactional data quality consists of human and technology perspectives; the data quality issues are driven by the context in which data is used, especially for reliability and maintenance and manually input data, which are more prone to human errors. For transactional data, the data quality control is highly weighted toward the way that the data is collected and reported in the first place. Therefore, it is important to create an implementation procedure on data quality management to continuously monitor, assure and control the transactional data quality. More efforts should be put on preparing the proper data quality implementation plan well before the data ingestion.

5.1 OVERVIEW OF THE DATA QUALITY MANAGEMENT PROCESS

This data quality management process is established in line with the best practices and standards as provided by ISO 8000-61 and ISO 14224. It proposes a framework to support systematic planning and implementation of data quality assessment, assurance and control in the marine and offshore industry. There are four phases, including:

- The "Plan" phase – The data quality planning phase. It establishes the data quality implementation plan to ensure the delivered results in accordance with data requirements
- The "Do" phase – The data quality deployment and control phase. It executes the data quality implementation plans and deploys the data corrective actions to handle the encountered data issues
- The "Check" phase – The data quality assurance phase. It monitors and measures the data quality and process performance against the data requirements and reports the results to validate the efficiency of the corrective actions
- The "Act" phase (data quality improvement) – The data quality improvement phase. It takes actions to improve data quality management process performance and prevents the recurrence of data nonconformity

The key steps and output per each phase of the process are described in the following four subsections. Figure 7 illustrates the detailed structure of transactional data quality management phases and steps.

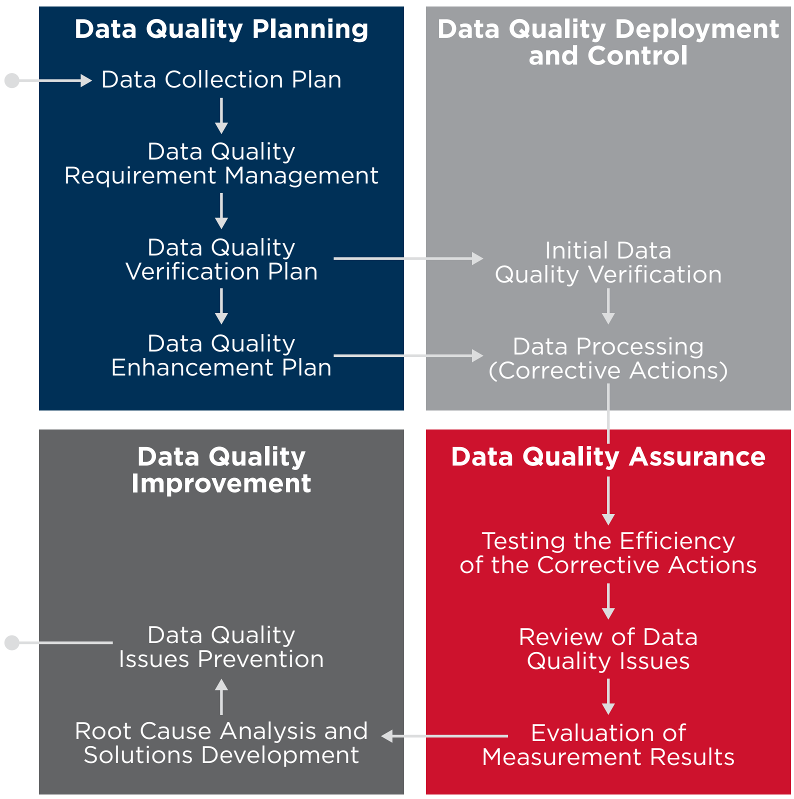


Figure 7: Transactional Data Quality Management Phases and Steps

5.2 DATA QUALITY PLANNING – "PLAN"

Confidence in the collected data, and hence any following analysis, is strongly dependent on the quality of the data collected. The data quality planning phase establishes data requirements and objectives for data quality based on the potential data use, creating implementation plans to achieve the objectives and the plan for evaluating the performance. These plans balance current data quality levels, costs, resources and capabilities across the organization for data quality management. This phase is initiated based on the needs and expectations of stakeholders or the feedback of the process improvements performed during data quality improvement (the act phase).

The data quality planning phase mainly consists of a data collection plan, data quality requirement management, a data quality verification plan and a data quality enhancement plan. All the plans mentioned below need to be reviewed and updated through consultation with stakeholders after a period of implementation.

5.2.1 Data Collection Plan

i. Describe the objective of data collection (e.g., the purpose of data application)

ii. Identify the data sources by using metadata, indicating the data mappings among the diverse data sources clearly. The sources need to be investigated to ensure that the sufficient data quality is provided (e.g., the inventory/technical equipment information is introduced, periodical sensor maintenance and calibration are scheduled; stocking replacement parts on site ensures that any part of the network can be replaced immediately)

iii. Identify the data collection boundary, indicating what data/parameters are to be collected

iv. Prepare data specifications by using metadata to identify data type, data format and data interpretations in accordance with the relevant international standards (e.g., ISO 14224, IMO compendium on facilitation and electronic business). Other supplementary information can also be provided, such as the ranges of data value and the level of details of data. The purpose of providing the data specification is to establish the basis for data processing and data quality monitoring and control

v. Identify data collection method and sequence based on a user guide or work instructions which can be provided by organizational data quality management specialists or equipment/machinery suppliers (e.g., how the data is collected especially when subsequent calculations are required)

vi. Identify a suitable structure for data partition and data storage for efficient data query

vii. Identify a standardized data collection platform or template for manual data input

5.2.2 Data Quality Requirement Management

i. Specify data quality requirements which are to be applied for performing data quality control, assurance and improvement through the data quality management process (ABS, 2019). The requirement definition may include data quality validation rules, corresponding measurable metrics and dimensions

ii. Define the acceptance level against each requirement (e.g., 85% of completeness of data is required)

iii. Prioritize the data quality requirements, which is based on a feasibility study of the requirements identified in terms of criticality of the potential data issue impact, technology, cost, timeliness and other severity measures

iv. Validate data quality requirements through the data quality management process, and when necessary, modify the requirements through consultation with stakeholders

5.2.3 Data Quality Verification Plan

i. In the "Plan" phase, assess the quality of a test data set as early as feasible in the data collection process according to the data quality requirements identified. The main objective of this early assessment is to look for any problems requiring immediate revision the planning measures to avoid unacceptable data. The early assessment also helps to test and select the data quality validation rules for performing data quality control and assurance in the "Do" and "Check" phases

ii. In the "Do" phase, evaluate the data quality levels of the actual data sets by implementing the data quality validation rules and measurement methods as identified in the plan phase. The purpose of data quality verification at this point is to propose data quality enhancement/corrective actions based on the measured data quality level

iii. In the "Check" phase, continue to monitor and evaluate the data quality levels of the remediated data sets against the identified data quality requirements. The purpose of data quality verification at this point is to validate the efficiency of the corrective actions. A time interval is set for periodically assessing the data quality against the defined validation rules

5.2.4 Data Quality Enhancement Plan

i. In the "Do" phase

- Diagnose data quality issues through an initial data quality check (as identified in Subsection 5.2.3/i) and discover the potential root causes of each data quality issue with assistance from the end users and subject matter experts. The purpose of root cause analysis is to identify the most efficient options for addressing data quality issues
- Identify the options for addressing both human and technical data quality issues. The focus is put on the technical root causes. For a short-term plan, the simple remediation such as the data cleansing and data correction can be performed. For a long-term plan, modifying the current data collection, transmission and/or storage process as well as improving the quality assurance and control to reduce human errors can be proposed to resolve root causes and prevent the issues from recurring, the detailed data quality improvement plan is introduced in the act phase

ii. In the "Check" phase, evaluate the data quality enhancement/remediation plan according to the quality level of the remediated data set. The purpose of this evaluation is to ensure the applied changes do not introduce additional errors and perform as expected

- Data quality risk assessment is required when the delivered data cannot meet the requirements in the corresponding data specification (e.g., data loss is very difficult or even impossible to recover). This assessment is used to establish the basis for further monitoring and control of processes through identifying the data quality risks, analyzing the potential impact and determining the risk priority

iii. In the "Act" phase, propose the improvement solutions to eliminate the root causes and prevent recurrence of nonconformities. Evaluate the feasibility of the proposed improvements through cost-benefit analysis

5.3 DATA QUALITY DEPLOYMENT AND CONTROL – "DO"

Data quality deployment and control phase executes all the data quality implementation plans established in data quality planning (see Section 5.2). This process involves assessing and processing data according to the specified verification and data enhancement plans.

Data quality deployment and control phase mainly consists of initial data quality verification and data processing (e.g., quality enhancement/corrective actions). All the processes mentioned below need to be reviewed and updated through consultation with stakeholders after a period of implementation.

5.3.1 Initial Data Quality Verification

The data quality level of the actual datasets is assessed initially at this step (as mentioned in Section 5.2.3/ii). The measurement indicators and corresponding metrics and dimensions are selected to measure the data quality level based on the defined data quality validation rules and the defined acceptance criteria. Both quantitative and qualitative manners can be used to measure the common data dimensions, including completeness, consistency, validity, timeliness and accuracy (ABS, 2019). The data quality validation rules defined in Section 5.2.2 should be tested and refined against the actual dataset with consideration of the encountered data problems and data use. Continuous refinement of the validation rules is required throughout the data quality improvement life cycle. Data stakeholders including end users and subject matter experts should be involved in this process. If data nonconformities are found, the simple remediation is to correct the data.

5.3.2 Data Processing (Enhancement/Corrective Actions)

The purpose of data processing is to deliver data that meet the requirements and work instructions in the corresponding data specification. By discovering the potential root causes of the data nonconformities with assistance from the stakeholders, data processing is deployed including the short-term data quality enhancement/corrective actions as defined in Section 5.2.4 (e.g., data cleansing, data parsing and formatting) to fix and update data in records. All the actions taken to address data nonconformities should be documented.

5.4 DATA QUALITY ASSURANCE – "CHECK"

The data quality assurance phase measures the data quality level and the process performance related to data nonconformities or issues that arise from data quality planning or data quality control. This measurement provides evidence by which to evaluate the impact of any identified poor levels of data quality on the effectiveness and efficiency of business processes. This process involves continuously monitoring and evaluating whether the enhanced data conforms to predetermined specification. If data nonconformities still exist, the remediation of root causes need to be further investigated.

The data quality assurance phase mainly consists of testing the efficiency of the corrective actions, review of data quality issues and evaluation of measurement results.

5.4.1 Testing the Efficiency of the Corrective Actions

Monitoring and measuring data processing performance and conformity of data can be performed at intervals or continuously in accordance with the data verification plan as predefined in Section 5.2.3. The purpose of data quality monitoring and measuring is to evaluate the efficiency of the short-term corrective actions, and then to identify the potential long-term remediation of root causes when data processing fails to deliver quality data that meet the requirements in the corresponding data specification. The options for addressing both human and technical related root causes are provided. The completeness of data mapping among the diverse data sources needs to be monitored continuously. If data nonconformities are still found, a record of the viability and degree of success for each corrective action needs to be prepared and distributed to stakeholders.

5.4.2 Review of Data Quality Issues

Data quality assurance is initiated in response to the unresolved data quality issues resulting from data quality planning and control. This review of data quality issues forms the basis for further investigations to improve data quality, including trends and patterns analysis in the occurrence of data nonconformities, the causes of stakeholder needs not being met and the ways in which an individual nonconformity can propagate to cause other nonconformities.

5.4.3 Evaluation of Measurement Results

The purpose of evaluation of measurement results is to establish the priorities for performing data quality improvement. The measurement results are presented in a quantitative manner. The impact of the identified poor levels of data quality or poor process performance needs to be investigated. In this phase, data quality risk assessment is performed to evaluate how potential data quality issues impact data analytics, operations, automation, reporting and decisions. The data quality-related risks should be identified throughout the data life cycle, ranked and managed by technical and business SMEs according to the probability and consequences of data quality related incidents defined in context.

5.5 DATA QUALITY IMPROVEMENT – "ACT"

The data quality improvement phase involves analyzing the root causes of data quality issues based on the assessment results derived from data quality assurance (see the "Check" phase). To prevent future data nonconformities, data quality improvement can be achieved by direct data correction (e.g., data cleansing and data enrichment) and prevention of the recurrence through improvement of the current data quality management processes.

The data quality improvement phase consists of root cause analysis and process improvement for data nonconformity prevention.

5.5.1 Root Cause Analysis

The root causes and associated impacts are analyzed for each identified data quality issue based on the results from the data quality assurance (see Section 5.4.2). The improvement solutions are proposed to prevent recurrence of identified root causes. The cost-benefit analysis is performed for each identified solution and the priority of the solutions is determined. Lastly, an improvement plan is proposed to eliminate the root causes and prevent recurrence of nonconformities.

5.5.2 Process Improvement for Data Non-conformity Prevention

By considering the improvement solutions as proposed through root cause analysis, a systematic approach to achieve data quality from an organization's perspective is established at this step.

A detailed proposal for improving the existing data management processes or suggestions of planned future processes is produced. A schedule is agreed with stakeholders for implementation of the process improvements. Once the agreed schedule is carried out, the effectiveness and efficiency of the process improvements can be evaluated by comparison to the situation before the implementation, for instance the extent to which data conformities are reduced or the extent to which the required resources can be measured.

6 ABS ROLE

As a classification society, the ABS Preventative Maintenance Program (PMP) consists of planned maintenance and/or condition-monitoring plans. Maintenance internals and tasks follow OEM recommendations or documented operator experience. It helps owners and operators maintain their vessels with updated machinery maintenance practices. ABS surveyors can then utilize the results of these programs to provide crediting toward the Special Periodical Survey Machinery requirements (CMS). This can reduce the amount of covered equipment being opened and minimize impact to operations from ABS survey activities. The ABS Preventative Maintenance Program requirements are contained in Section 7-A-14 of the ABS *Rules for Survey After Construction (Part 7)*.

ABS has published the *Guidance Notes on The Implementation of Smart Functions Onboard Marine and Offshore Vessels* (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. These Guidance Notes provide 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).

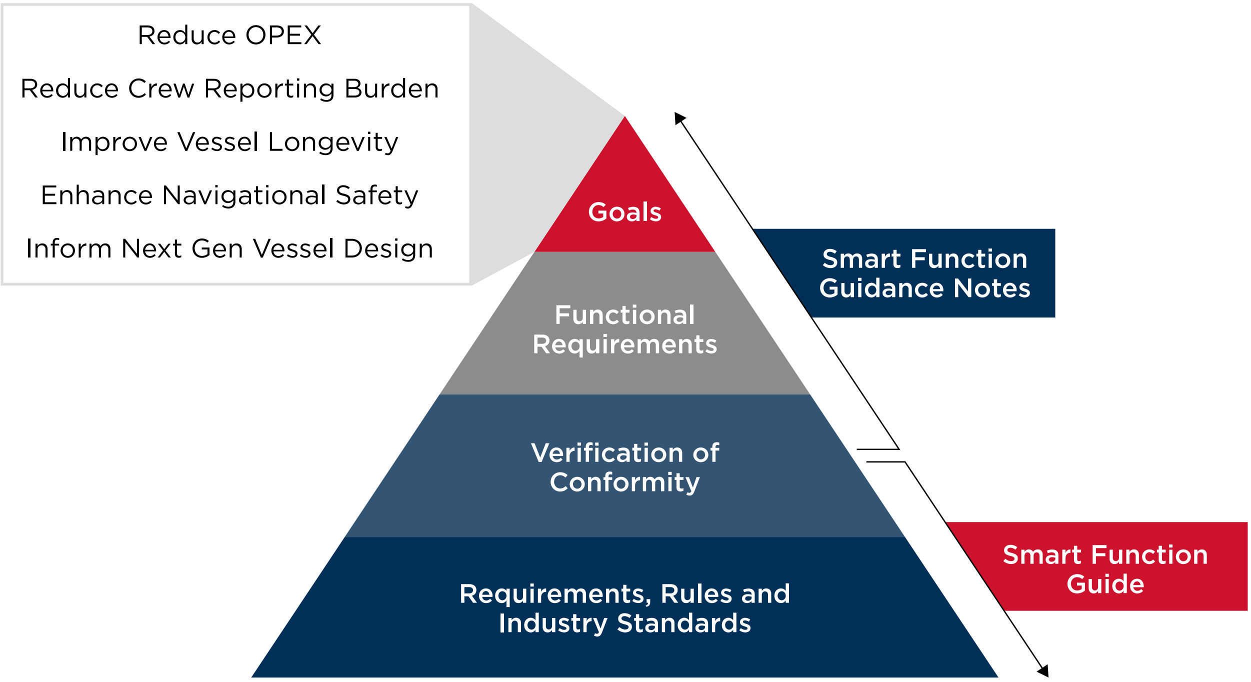


Figure 8: Goal-based Smart Function Framework

ABS has also published the Guide for Smart Functions for Marine Vessels and Offshore Units (the Smart Guide) to define the risk-informed technical requirements for independently assessing the Smart Function system(s) and offer optional SMART class notations and Class Record Comments (refer to Figure 8 for the coverage of the Guidance Notes and The Guide).

ABS has launched the ABS My Digital Fleet™, a state-of-the-art digital platform for marine and offshore operators to connect their data and help improve fleet efficiency, reduce costs and manage risks. My Digital Fleet™ is a configurable intelligence platform that can interface with client systems and provides real-time, data-driven insights. Users can easily understand an asset's performance in terms of regulatory compliance, fuel efficiency, structural and mechanical integrity.

As a means to credit Continuous Survey, ABS needs to verify the equipment maintenance records for vessels enrolled in the Preventative Maintenance Program (PMP), and CMMS data can be used to verify regulatory compliance. Moreover, both the smart functions, as defined the Smart Guidance Notes and Smart Guide and My Digital Fleet™ rely on quality data and data analytics to derive health and performance conditions, provide operational insights and assist in decision-making Data quality assessment, assurance and control is a key component for such data-driven implementation. ABS, as a classification organization, 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. In addition, ABS, as a trusted partner, will work with the industry to promote a data quality culture, mature data quality management and improve the overall data quality for more reliable data-driven applications.

7 SUMMARY

The marine and offshore industry has entered an era of big data and accelerated digital transformation, and managing transactional data is an important aspect of a data-driven analytics journey. High quality data is the essential ingredient for high-quality analytics. Unfortunately, the maritime industry has limited experience on systematic planning and implementation of data quality assessment, assurance and control to facilitate the data analytics. Data quality assessment and improvement require not only in-depth understanding of the data and data quality but domain support as well.

To tackle the above issues, this Advisory has introduced an adaptable data quality management process built upon ISO 8000-61 and ISO 14224. It provides a holistic approach for managing transactional data in four phases, including: i) the preparation of data quality implementation plan for standardization of data collection exercises before the data collection process starts; ii) execution of data quality implementation plans for identifying and addressing data quality issues through rigorous data profiling, data processing and control of incoming data; iii) testing the efficiency of the corrective actions, continuously monitoring and measuring the data quality levels; and iv) performing root causes analysis for unresolved data quality issues and proposing the improvement plan for data quality issues prevention.

Unlike the time-series data, the quality of transactional data is affected by both human factors and technology limitations and is highly dependent upon the standardization of the data collection and reporting process. Similar data quality dimensions of the time-series data may be applied to transactional data. However, it is typically hard to use quantitative formulae to define the quality indicators (e.g., data quality metrics) for transactional data, as the number of observations (e.g., the number of data quality check failures) or the total sample size can vary with different operation conditions and the lack of fixed data frequency.

The study on marine transactional data quality management is still in early stages of development. More industry focus should be put on the use cases to improve the applicability of the current best practices of data quality assessment, assurance and control during operation.

8 APPENDIX 1 - COMMON DATA QUALITY PROBLEMS WITH RELIABILITY AND MAINTENANCE DATA

This appendix intends to map the typical reliability and maintenance data quality issues to the existing data quality dimensions for the purpose of data quality assessment. Moreover, each issue is also linked with the critical data management phases/steps as defined in ISO 8000-61, to provide recommendations on data quality assurance and control.

Data Quality Issues	Data Quality Dimensions	ISO 8000-8 Data Quality Category	Management Stage - "PDCA" Cycle	Recommendations on Data Quality Assurance and Control
Insufficient sample size - Small sample sizes of the failure data give improper results for the failure probability/reliability distribution functions	Completeness; Usefulness	Pragmatic	"Do" and "Act"	• "Do" – Initial data quality verification to check if the collected data gives acceptable statistical confidence; Define data quality enhancement/correction method for a particular case • "Act" - Provide feedback to data quality enhancement/correction plan (e.g., increase the sample size and start a new iteration). It needs to evaluate the impact of sample size on standard error of the mean (SEM); SEM is reduced with increasing sample size
Wrong information on TTF (time to failure) – The time to failure information is not accessible since CMMS was installed after the equipment/system start-up. Lots of failure events are not recorded in the system. The reliability is overestimated	Accuracy; Accessibility; Usefulness	Pragmatic	"Do" and "Act"	• "Do" – Initial data quality verification to check if the collected data complies with definition and interpretation rules; Define data quality enhancement/ correction method for a particular case • "Act" - Provide feedback to data quality enhancement/correction plan (e.g., a correction method for TTF computation is to be provided)
Inconsistency - It is often caused by merging different data sources; such as inconsistencies in serial number (i.e., unit identifiers), product type code (i.e., the product identified), fault codes, parameter units and data format (e.g., the failure data is reported as a date)	Consistency	Syntactic	"Plan" and "Do"	• "Plan" - Define meta data specifications to express data format required for data processing • "Do" - Initial data quality verification (e.g., data profiling tool) to check the inconsistency values according to the desired data format

Logical errors – Failure events referring to an asset or functional location which does not exist; Functional locations referring to equipment which does not exist	Validity; Completeness	Semantic	“Plan” and “Check”	• “Plan” – Provide data specifications to identify what equipment data, failure data and maintenance data are to be collected • “Check” – Review the completeness of data mapping (e.g., failure events vs. locations, equipment vs. locations)
Illogical order of dates (e.g., failure date occurs before corresponding reporting date or manufacturing date)	Validity; Timeliness	Syntactic; Semantic	“Plan” and “Do”	• “Plan” – Define meta data specifications to identify the crucial dates for data quality verification • “Do” - Initial data quality verification (e.g., data profiling tool) to check the order of dates according to the defined timelines
Missing Rows - Missing entries about products or failure events, which affect the calculation of mean time to failure; The reliability is overestimated Missing columns in an entire row - e.g., the serial number, operation start date and failure date are essential for reliability calculation. Running hours and disposal date can provide additional information that are useful for an accurate analysis	Completeness (Row) Completeness (Column)	Syntactic; Semantic	“Plan,” “Do,” “Check” and “Act”	• “Plan” – Define the acceptance levels on data completeness; Define the specific data quality rules to calculate the percentage of missing rows and columns • “Do” – Initial data quality verification to check the current completeness of entire dataset; Define data enrichment method to fill in the missing values if it is necessary. It may reflect the assumptions and distort the reality. The data enrichment/ correction needs to be recorded for monitoring or audit • “Check” – The data collection procedure needs to be monitored and periodically reviewed. Testing the efficiency of the data control methods (enrichment/corrections) • “Act” – Review the data issues and provide feedback on planning stage
Reliability data is artificially changed by service technicians or SEMs based on their domain judgment; some implausible data may represent a valid outlier	Plausibility	Semantic	“Do”	• “Do” - Clearly identify the data processing method for corrective actions. The data correction needs to be recorded for monitoring or audit

Low in richness of information (e.g., use unstructured 'free-text' failure description to a software (i.e., CMMS) rather than a standardized failure code; it will affect the accuracy of the advanced reliability analysis, or use week or month instead of specific failure dates, etc.)	Accuracy	Semantic	“Plan” and “Do”	• “Plan” – Define meta data specifications to express data format required for data processing and data analytics • “Do” - Initial data quality verification (e.g., data profiling tool) to check the level of details of the data
Manual data-entry errors - cause a wrong quantity request for a material. The current stock cannot fulfill the request of replenishment order (e.g., order 1,000 units instead of 100). Manual data errors may result in wasted capital	Accuracy	Pragmatic	“Plan”	• “Plan” - Get a proactive Enterprise Resource Planning (ERP) system which offers time-saving automation. Main elements within the automated system for spare parts control are periodic review (e.g., daily, weekly, or monthly), demand forecasting, lot sizing and safety stocks. The stock level can be controlled based on predicted demand within lead times
Improper equipment registers and lack of real-time track. More databases contained information on equipment, thus changes were not always properly registered in all systems (e.g., the equipment part is no longer in production)	Currency; Accuracy	Pragmatic	“Plan”	• “Plan” - For the slow-moving spare parts (slow movers), a database is required for recording the obsolete spares in the market. Substantial costs and long lead time are involved for manufacturers in setting up a production run especially for the obsolete part

Lack of the failure data to determine a spare part demand pattern during the use of the newly issued equipment. The slow movers may take a long time before a good demand history is built up. The intermittent high demands may mislead the parts are fast movers.	Completeness	Semantic	“Check”	• “Check” - Demand patterns may fluctuate in time. The estimated demand rates need to be periodically reviewed and frequently updated
Wrong classification on equipment criticality (risk level). This is a subjective process. All the parts needed for critical equipment should have a higher service level and higher downtime penalty costs than parts in low critical equipment	Accuracy	Semantic	“Plan” and “Check”	• “Plan” - Identify a risk-based or critical stock levels for the spare parts stock control • “Check” - A periodic review and update for the ordered items and the remaining stock level are required

9 APPENDIX 2 - COMMON DATA QUALITY PROBLEMS WITH MANUAL INPUT DATA

This appendix intends to map the typical manual input data quality issues to the existing data quality dimensions (ABS, 2019), data quality categories (ISO 8000-8:2015, Nov, 2015) and the critical data management phases (ISO 8000-61:2016, 2016) for continuous data quality assurance and control.

Data Quality Issues	Data Quality Dimensions	ISO 8000-8 Data Quality Category	Management Stage - "PDCA" Cycle	Recommendations on Data Quality Assurance and Control
The record values for the specific period do not reflect the reporting period. Some of them are instantaneous, direct- recorded from sensors	Accuracy	Semantic	"Plan"	• "Plan" - Define what objective of the collected data is and what application the data is expected for. Investigate the source(s) of the data to ensure that adequate data quality is provided.
Data recorded based on a subsequent calculation rather than a direct measurement. The calculation is not performed according to the recommended procedure	Accuracy	Semantic	"Plan" and "Check"	• "Plan" - Prepare data collection plan and clearly define the data collection sequence. The ship's crew needs to follow the guide/the exact procedure provided by the engine maker/vendor to do proper calculations • "Check" - The data collection procedure needs to be periodically reviewed
Subjective judgement/forced value – Values are estimated by crew with subjective judgement, such as the wave conditions	Accuracy; Plausibility	Semantic	"Plan," "Check" and "Act"	• "Plan" - Provide data specifications to identify the data collection process and define suitable data collection methods • "Check" – Review the data collection procedure periodically • "Act" – Review the data issues and provide feedbacks on planning stage.
Data format/ pattern – Different data format/ pattern for the same data element (e.g., different date/time format). Specifically, when input is free text, the location and failure are described using different name or format of names. It makes data manipulation/ processing difficult	Consistency	Syntactic	"Plan" and "Do"	• "Plan" - Define meta data specifications to express data format required for data processing • "Do" - Initial data quality verification (e.g., data profiling tool) to check the inconsistency values according to the desired data format

Ambiguity – Different level of data accuracy for the same data element (e.g., a precision of five (5) decimal places allows different functionalities rather than a precision of two (2) decimal places)	Precision	Syntactic	"Plan" and "Do"	• "Plan" – Define meta data specifications to specify the level of data accuracy (e.g., significant digits) • "Do" - Initial data quality verification (e.g., using data profiling) to check the data values according to the defined level of data accuracy
Missing events – Missing rows and missing columns	Accuracy; Completeness	Syntactic; Semantic	"Plan," "Do," "Check" and "Act"	• "Plan" – Define the acceptance levels on data completeness • "Do" – Initial data quality verification to check the current completeness of entire dataset; Define data enrichment method to fill in the missing values if it is necessary. It may reflect the assumptions and distort the reality. The data enrichment/ correction needs to be recorded for monitoring or audit • "Check" – The data collection procedure needs to be monitored and periodically reviewed. Testing the efficiency of the data control methods (enrichment/corrections) • "Act" – Review the data issues and provide feedback on planning stage
Values out of range – Data values are out of range for the domain under observation (e.g., value spikes or sudden changes which are implausible)	Validity	Syntactic; Semantic	"Plan" and "Do"	• "Plan" – Define meta data specifications to assign the value range within a defined numeric, lexicographic or time range • "Do" - Initial data quality verification (e.g., using data profiling) to check the data values according to the defined data range. Define the data correction method for data processing
Wrong timestamps or wrong timestamp order – Timestamps are mismatched to the expected time (e.g., time gaps/ redundancies exist). Timestamps are not in chronological order	Timeliness	Syntactic	"Plan" and "Do"	• "Plan" – Define meta data specifications to identify the reporting dates/periods for data quality verification • "Do" - Initial data quality verification (e.g., data profiling tool) to check the order of dates according to the defined timelines

A sentence in the inspection and survey report is constructed poorly (spelling, grammar); gives completely different meaning from what the surveyor wants to express	Accuracy	Semantic	"Plan" and "Check"	• "Plan" – Prepare data collection plan and establish a standard data reporting template/platform to limit the subject input (e.g., selection from a drop-down list) • "Check" - All reports need a careful proofing by a professional surveyor or the third party if the attending surveyor feels he/she is not competent in a particular area
Low in richness of information - The digital photographs are widely used for marine survey reports over the past 10 years. In some circumstances, surveyors only use captions adhered to the pictures instead of formal reports; lack of the supporting texts/ sentences to explain the points that he/ she wants to make	Completeness;	Semantic	"Plan" and "Check"	• "Plan" – Define meta data specifications to define the level of details of the survey information to be reported; Create a suitable structure to partition and store the vast number and variety of survey reports for fast and efficient data query • "Check" - The data collection and reporting procedure needs to be periodically reviewed

10 APPENDIX 3 - COMMON DATA QUALITY PROBLEMS WITH SENSOR DATA

This appendix intends to map the typical sensor data quality issues to the existing data quality dimensions (ABS, 2019), data quality categories (ISO 8000-8:2015, Nov, 2015) and the critical data management phases (ISO 8000-61:2016, 2016) for continuous data quality assurance and control.

Data Quality Issues	Data Quality Dimensions	ISO 8000-8 Data Quality Category	Management Stage - "PDCA" Cycle	Recommendations on Data Quality Assurance and Control
Missing data - The sensor stops providing any reading on its interface	Accuracy; Completeness	Syntactic; Semantic	"Plan," "Do," "Check" and "Act"	• "Plan" – Define the acceptance levels on data completeness. Take a complete backup of all essential data to prevent data disruption • "Do" – Initial data quality verification to check the current completeness of entire dataset; Define data enrichment method to fill in the missing values if it is necessary. It may reflect the assumptions and distort the reality. The data enrichment/correction needs to be recorded for monitoring or audit • "Check" – The data collection procedure needs to be monitored and periodically reviewed. Testing the efficiency of the data control methods (enrichment/corrections) • "Act" – Review the data issues and provide feedback on planning stage
Stuck/jammed data – The sensor reading gets jammed and stuck in some incorrect value	Accuracy; Currency	Semantic	"Plan" and "Do"	• "Plan" – Define the data collection plan to test and calibrate the sensors periodically. The sensor baseline performance can be provided • "Do" - Initial data quality verification (e.g., using data profiling) to check the stuck data values. Define the data enhancement/correction methods for a particular case
Bias - Constant offset or temporal delay; the observations continuously deviated from the expected value by a constant offset, or continuously produced with a constant temporal deviation	Accuracy	Semantic	"Plan" and "Do"	• "Plan" – Define the data collection plan to test and calibrate the sensors periodically. The sensor baseline performance can be provided • "Do" - Initial data quality verification (e.g., using data profiling) to check the biased data values. Define the data enhancement/correction methods for a particular case

Invalid data type – Invalid data type for the same data element (e.g., a text string found in a list of floating numbers)	Structural Validity	Syntactic	“Plan” and “Do”	• “Plan” – Define meta data specifications to identify data type required for data processing (e.g., text, number or date/time); Take a complete backup of all essential data to prevent data disruption • “Do” - Initial data quality verification (e.g., data profiling tool) to check the structural invalidity values according to the meta data specification
Different data format/pattern – Different data format/pattern for the same data element (e.g., different format in a timestamp that makes data manipulation/ processing difficult)	Structural Consistency	Syntactic	“Plan” and “Do”	• “Plan” - Define meta data specifications to express data format required for data processing • “Do” - Initial data quality verification (e.g., data profiling tool) to check the inconsistency values according to the desired data format
Wrong timestamps or wrong timestamp order – Timestamps are mismatched to the expected time (e.g. time gaps/ redundancies exist). Timestamps are not in chronological order	Timeliness	Syntactic	“Plan” and “Do”	• “Plan” – Define meta data specifications to identify the reporting dates/periods for data quality verification • “Do” - Initial data quality verification (e.g., data profiling tool) to check the order of dates according to the defined timelines

11 APPENDIX 4 - DEFINITIONS

Data: Reinterpretable representation of information in a formalized manner suitable for communication, interpretation or processing [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 [Source: ISO/IEC 11179-3:1994].

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

Data Life Cycle: Stages in the management of a data [Source: ISO/IEC 20547-3:2020].

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 represents the measurable feature or characteristic of data. Each dimension consists of a single or a set of data quality metrics. [Source: ISO 8000-8, 2015].

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

Data Quality Metrics: A data quality metrics is an indicator that represents the data quality performance of data records (sensors) in a data set. [Source: ISO 8000-8, 2015].

Data Quality Validation Rule: A data quality rule is defined to describe data quality requirements for determining/validating conformance of data to expectations. [Source: ISO 8000-8, 2015].

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 20924:2018].

Metadata: Data that defines and describes other data [Source: ISO/IEC 11179-1:2004].

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].

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

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

Special Continuous Survey of Machinery (CMS): A program in which the total number of survey items is arranged in order to provide for survey of approximately 20 percent of the machinery each year during a five-year period. [Source: Section 7-A-14 of the ABS Rules for Survey after Construction (Part 7)].

Time-series Data: Sequence of data values which are ordered in time. [Source: ISO 19848, 2018].

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

Verification: Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled [Source: ISO 9000:2015].

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].

12 APPENDIX 5 - REFERENCES

ABS. (2018). *Guidance Notes on Smart Function Implementation.*

ABS. (2019). *Advisory on Data Quality for Marine and Offshore Application - Time Series Data.*

ABS. (2019). *Guide for Smart Functions for Marine Vessels and Offshore Units.*

ABS. (2017). *Rules for Survey After Construction.*

Ali Akbar Safaei, Hassan Ghassemi, Mahmoud Ghiasi. (2018). Correcting and Enriching Vessel's Noon Report Data. *European Transport \ Trasporti Europei (2018) Issue 67, Paper n° 7, ISSN 1825-3997.*

ASQ. (2019, Oct.). *What is the Plan-Do-Check-Act (PDCA) cycle?* Retrieved from American Society for Quality: https://asq.org/quality-resources/pdca-cycle.

Dongmei Huang, Danfeng Zhao, Lifei Wei, Zhenhua Wang, Yanling Du. (2015). Modeling and Analysis in Marine Big Data: Advances and Challenges. *Mathematical Problems in Engineering, Hindawi.*

Hui Yie Teh, Andreas W. Kempa-Liehr, Kevin I-Kai Wang. (2020). Sensor data quality: a systematic review. *Journal of Big Data.*

Informatica. (2020, Oct. 15). Retrieved from What is Unstructured Data? https://www.informatica.com/sg/services-and-training/glossary-of-terms/unstructured-data-definition.html.

Informatica. (2020, Oct. 15). *IoT Data Management: The Rise of Industrial IoT and Machine Learning.* Retrieved from https://www.informatica.com/sg/resources/articles/iot-data-management-and-industrial-iot.html

ISO 14224:2016. (2016). *Petroleum, petrochemical and natural gas industries - Collection and exchange of reliability and maintenance data for equipment.* Retrieved May 21, 2018, from https://www.iso.org/standard/64076.html.

ISO 8000-61:2016. (2016). *Data quality - Part 61: Data quality management: Process reference model.*

ISO 8000-8:2015. (Nov, 2015). *Data quality - Part 8: Information and data quality: Concepts and measuring.*

M. R. Hodkiewicz, N. Montgomery. (2014). Data fitness for purpose: assessing the quality of industrial data for use in mathematical models. *8th IMA International Conference on Modelling in Industrial Maintenance and Reliability.* Oxford.

Ralf Gitzel, Simone Turring, Sylvia Maczey. (2015). A Data Quality Dashboard for Reliability Data. *2015 IEEE 17th Conference on Business Informatics*, (pp. 90-97). Lisbon, Portugal.

Ralf Gitzel, Subanatarajan Subbiah, Christopher Ganz. (2018). A Data Quality Dashboard for CMMS Data. ICORES 2018 - *7th International Conference on Operations Research and Enterprise Systems*, (pp. 170-177). Portugal.

Ricardo Pérez-Castillo, Ana G. Carretero, Ismael Caballero, Moisés Rodríguez, Mario Piattini, Alejandro Maté, Sunho Kim, Dongwoo Lee. (Sep. 2018). DAQUA-MASS: An ISO 8000-61 based data quality management methodology for sensor data. *MDPI.*

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

AFRICA AND 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
World Trade Tower, 29F, Room 2906
500 Guangdong Road, Huangpu District,
Shanghai, China 200000
Tel: +86 21 23270888
Email: ABSGreaterChina@eagle.org

NORTH PACIFIC REGION
11th Floor, Kyobo Life Insurance Bldg.
7, Chungjang-daero, Jung-Gu
Busan 48939, Republic of Korea
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 05/21 21099