2025 IMEKO TC-6 International Conference on
Metrology and Digital Transformation - **M4DConf 2025**
Benevento, Italy, September 3-5, 2025

Enhancing digital twin reliability using FAIR principles and data quality assessment

Miguel Burg Demay[1], Luiz Eduardo de Farias[2], Gustavo Donatelli[3], Andre Luiz Meira de Oliveira[4]

[1] *Certi/Centro de Metrologia e Instrumentação, 88040-535, Florianópolis/Brazil, mbd@certi.org.br*
[2] *Certi/Centro de Metrologia e Instrumentação, 88040-535, Florianópolis/Brazil, lzr@certi.org.br*
[3] *Certi/Centro de Metrologia e Instrumentação, 88040-535, Florianópolis/Brazil, gd@certi.org.br*
[4] *Certi/Centro de Metrologia e Instrumentação, 88040-535, Florianópolis/Brazil, aeo@certi.org.br*

***Abstract* – The use of digital tools such as digital twins is spreading throughout the O&G sector. For integrity management, digital models of relevant asset degradation phenomena have been used to estimate and predict its health, aiming to improve maintenance planning and to provide information about the risk of failure and its evolution over time. The input data of models for O&G integrity monitoring are commonly found in different databases and present very different characteristics, such as sampling rate, temporal stability and variability, influencing the data quality in different ways. This work addresses the use of FAIR principles and data quality assessment for O&G asset integrity management. A brief review of established concepts is discussed, and a practical case study is presented, which illustrates the very important role that data quality assessment and the use of FAIR principles play in digital models' reliability.**

I. INTRODUCTION

The integrity management of subsea assets has traditionally depended on inspections performed by tools such as pigs and ROVs. While these methods provide accurate measurements at specific locations, the entire process, from planning to execution, is both costly and time-consuming, resulting in less frequent monitoring than would be ideal. Furthermore, assets design limitations can prevent inspection tools from accessing the most critical areas, which need the most attention.

Digital solutions provide a powerful alternative to overcome these inherent limitations. They involve creating digital models that simulate the key degradation processes impacting subsea assets. When integrated into a digital twin, these models enable continuous integrity monitoring. By analyzing real-time production data and fluid properties, the models can evaluate an asset's health and its future degradation at a significantly higher frequency and lower cost than traditional methods, though with greater uncertainty [1][2].

The reliability of an asset degradation model's estimates or predictions relies heavily on the quality of the input data. In the oil and gas industry, the data used for these models come from various sources and have distinct characteristics, such as sampling rate and temporal variability. In addition, gaps often appear on the data due to sensor malfunctions and variables with different sampling rates within the same dataset.

This paper explores how the application of FAIR principles and the evaluation of data quality can be beneficial in this context. Section II introduces the concept of a digital twin for oil and gas asset integrity management. Sections III and IV discuss the FAIR principles for measurement data and data quality assessment. Sections V and VI illustrate how the FAIR principles and data quality assessment can be applied for a digital twin for O&G integrity management. Finally, Section VII presents the main conclusions of this work.

II. DIGITAL TWIN FOR SUBSEA ASSET INTEGRITY MONITORING

A digital twin (DT) can be described as a set of digital models representing a physical asset, designed to simulate and predict its current and future states and behaviors, while enabling control through feedback to the physical asset. Data and digital models form the foundation of this concept. [3][4]

As shown in Figure 1, data and information serve as the link between the physical and digital worlds. In the physical domain, materials, machines, operators, and the environment are key elements in the production process.

Otherwise, in the digital domain, various models representing different aspects of the production process support accurate predictions and informed decision-making. These models not only describe asset behaviors, but also utilize historical and real-time data, as well as human experience and knowledge, to predict the future conditions of assets [5]

To ensure good performance, synchronized behavior is expected from both physical and digital twins. Through simulations, the digital model optimizes operations and improves the performance of the physical process by offering feedback. This ongoing enhancement requires the co-evolution of both the physical asset and its digital twin [6][7][8].

Fig. 1. Physical and digital interaction in Digital Twin.

Developing a digital twin management system involves integrating and evaluating the quality of data used by the degradation models associated with the asset's function. The digital twin helps assess the asset's health by monitoring key parameters for each degradation model, allowing for the identification of potential failure modes and supporting risk-based decision-making [2].

In the context of subsea assets, erosion is one of the most important degradation phenomena. It impacts the internal surfaces of pipelines, gradually reducing wall thickness and shortening their useful life. Several digital models have been developed to estimate the loss of pipeline wall thickness due to erosion, with one of the most widely used being the model presented in DNV-O501[9].

The DNV-O501 model relies on over 30 input parameters, sourced from multiple databases which use varying data formats and sampling rates. These data differ in characteristics such as temporal stability and variability, each influencing the model's output in distinct ways.

In this complex scenario, applying FAIR principles and assessing data quality could enhance the reliability of digital model outputs, as well as improve the accuracy of digital twin estimates, predictions, and recommendations.

III. FAIR PRINCIPLES

The FAIR principles are a set of guidelines intended to support the development of data management systems for data, metadata, and infrastructure. These principles aim to enhance data management and prepare data for sharing and reuse across stakeholders from different domains, disciplines, or organizations by ensuring that data is Findable, Accessible, Interoperable, and Reusable [9][10][11].

A findable dataset means that data (and metadata) should be easily discoverable by humans and machines. Machine-readable metadata are essential for the automatic discovery of datasets and services. For that, data should be described with comprehensive metadata, and both should be assigned to a globally unique and persistent identifier.

An accessible dataset is one where users know precisely how to access the data, including any necessary authentication and authorization. For that, data (and metadata) should be retrievable by means of an open, free and standardized communication protocol.

Interoperability refers to the requirement that data must be compatible and integrated with other data. Additionally, the data must be able to interact with applications or workflows for analysis, storage and processing. For that, data and metadata should use a formal, accessible, shared and widely applicable language for knowledge representation.

Data reuse aims to maximize the value of data by enabling its reuse. To accomplish it, data and metadata should be well-described to ensure they can be reliably replicated and/or integrated across different contexts.

By following these four principles, datasets become accessible to a broader range of data users, allowing them to easily find, access, understand, and automatically process the data.

The interest in aligning with the FAIR principles has been increasing since they ease data integration, which is essential for linking and combining data across various fields. Adhering to the FAIR enhances the value of data, by making it more discoverable through unique identifiers and more easily integrable through standardized and shared knowledge representations [11][12].

IV. DATA QUALITY

Data is essential in the industry for monitoring, controlling and optimizing production processes, as well as for conducting analyses and simulations that support decision-making. To ensure the results meet the necessary levels of reliability, the data and models must meet distinct quality criteria.

Data quality has been gaining increasing relevance in the oil and gas industry, being addressed in at least two recommendations, DNV-RP-0497 and DNV-RP-A204.

Data quality can be understood as the degree to which the data is suitable for the analysis that is intended to be performed. Therefore, the quality of data cannot be assessed without considering the data's intended application.

When it comes to digital twins, which rely on simulations and data analysis, the quality of input data is crucial for obtaining useful and reliable results. To build trust with the user and contribute to the company's business, a digital twin must operate with valid and accurate data, along with representative models. High data quality must be ensured to enable data reuse and the analysis of historical data.

To measure data quality, different metrics can be used, often grouped into dimensions of data quality. These dimensions can also be found in ISO 8000-8, which describes data and information quality concepts and their importance for quality management processes. Among the various existing metrics, interoperability and accessibility stand out, along with [13][14][15]:

- Accuracy: a measure of how much data represent the real-world values
- Completeness: the extent to which data is available.
- Consistency: related to the uniformity of data across different databases and systems.

- Timeliness: a measure of how much data is up-to-date and available

V. FAIR PRINCIPLES AND DATA QUALITY FOR O&G DIGITAL TWINS

Digital twins for integrity management in the oil and gas industry usually have several digital models for estimating asset degradation, which consume data of different types and characteristics, scattered across various databases.

Production data, like pressure and temperature, are acquired by sensors, with a sampling rate of the order of tens of hertz. Fluid characteristic data, such as density and viscosity, are obtained through laboratory tests and are collected monthly, bimonthly, or semi-annually. Asset specification data, such as its dimensions and materials, remain constant over time.

To ensure that data adhere to the FAIR principles, one approach is to build an unified database, from which all digital models can draw the data they need. This database shall mean the only digital truth of the physical twin.

Nonetheless, data read from different databases, formatted, and stored in a single database may present some gaps due to the different sampling frequencies and invalid values. Therefore, issues related to the validity of the data in this dataset must be addressed, to ensure that the models always have valid data to operate.

This digital twin's unified database should include metadata, which contain guiding information for data conditioning, such as:

- Type of data: original, conditioned, or synthetic
- Measurement range of each variable
- Physical limits of each variable
- Data quality indicators

The conditioning process involves removing invalid data and filling in gaps, ensuring alignment with the FAIR principles. To realize it, gaps can be filled, for example, with the most recent valid data. The same approach can be applied to data that are not realistic, such as negative pressures or excessively high temperatures.

Thus, conditioning represents a decrease in data quality, as it involves inserting a value based on the best available knowledge, which may not reflect reality. Therefore, each conditioning operation reduces data quality, but it is better than using invalid values.

Additionally, when data is replaced by the most recent valid value, larger gaps are filled with the same value. As a result, the data quality diminishes over time or with the number of sequential replacements, since the time between the replaced data and the most recent valid value increases, thereby reducing its representativeness.

In this sense, it is very important to understand the deep relationship between data quality and uncertainty, and model sensitivity.

The output of a digital model contains uncertainties that arise from the model input data and the modelling process. The former, called aleatoric uncertainty, is related to measurement uncertainties, process and fluid variations, which compose the natural variability of input data and should not impact its data quality. However, outliers and unknown variations should be considered when assessing data quality.

The uncertainty related to the modelling process is named epistemic uncertainty. It is associated with the completeness of the knowledge represented by the model. In other words, it is related to the representativeness of the model in comparison to the real behavior of a degradation phenomenon. [16][17][18][19]22].

This kind of uncertainty is associated with input data quality through the sensitivity of each model parameter. Since each variable impacts differently the model outputs, the quality of input data influences the quality of output data according to its sensitivity [18][19].

VI. DATA QUALITY INDICATOR EVALUATION

Based on these premises, a quality indicator value can be defined for each data in the dataset. For each input data of the DNV-O501 erosion model considered in this paper, a quality indicator can be assigned, which is a measure of reliability of the data used as input in erosion models for subsea equipment.

Data quality indicator (DQI) can be set ranging from 0 to 1, where 0 represents a complete lack of quality and 1 represents the highest quality. To evaluate data quality indicators, four parameters have been considered:

- Data validity: a measure of the validity of each input data, which is decreased by the presence of null or invalid values. For each day, the data quality is calculated as the average of the daily valid records. If all the records are valid, the quality of the input data is 1; otherwise, the quality is reduced proportionally to the number of invalid records.
- Physically possible values: it checks if the data falls within a range of physically possible values for each variable. For example, the temperature of an underwater equipment cannot be lower than 0 °or higher than 100 °C. If the data is within the range, the quality is 1; otherwise, it is 0.
- Coefficient of variation: it evaluates the variation of the data over time. If it is greater than the historical median or equal to zero, the quality is reduced. This helps identify fluctuating or inconsistent data.
- Timeliness: it is also possible to include a quality degradation rate, which reduces the quality of the data over time if it is not corrected or updated. For example, if a sensor fails and is not repaired, the quality of the data recorded by that sensor will gradually decrease until it reaches a minimum value. This approach ensures that the data quality reflects not only its consistency at the time of collection but also its reliability over time.

Based on these parameters, each model's input data can have its quality quantitatively assessed. However, for management purposes, a qualitative approach can be interesting. For that, the classification system presented in table 1 can be used.

Table 1. Qualitative Data Quality Indicators

Quality class	Condition
High quality data	DQI ≥ 0.85
Intermediate quality data	0.5 ≤ DQI < 0.85
Low quality data	DQI <0 .5

Based on these definitions, a dataset of an O&G asset was assessed for degradation caused by erosion. For a 10-year time window, the DNV-O501 erosion model data inputs were recorded, prepared and stored in a database, as well as its DQI as metadata.

Figure 2 illustrates a DQI of a X-tree pressure signal over time. Daily values were obtained and represented in the graph, which shows how the quality of the pressure signal can vary over the years.

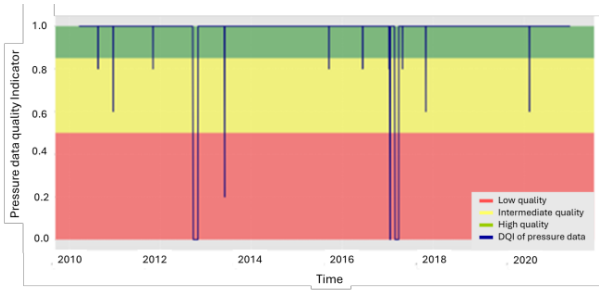

Fig. 2. Data Quality of pressure data

For each input variable, its DQI is evaluated through a weighted average of the daily quality assessments, where the weight of each rating (high, intermediate or low) corresponds to its frequency during the analyzed period.

Then, for each input data, a DQI could be evaluated and stored. Among all input variables of the DNV erosion model, nine relevant DQIs are presented in table 2.

Table 2. Data Quality of DNV-O501 model input data

Input data	DQI	Quality class
Oil rate	0.72	Intermediate quality data
Pressure	0.97	High quality data
Temperature	0.61	Intermediate quality data
GOR	0.72	Intermediate quality data
BSW	0.72	Intermediate quality data
Oil density	0.50	Intermediate quality data
Oil viscosity	0.50	Intermediate quality data
Gas viscosity	0.39	Low quality data
Sand content	0.50	Intermediate quality data

According to table 2, pressure has a high data quality; gas viscosity presents low data quality, and all the others have an intermediated data quality.

Based on this procedure, the quality of each model input data can be assessed. While this is important for asset's integrity management, the quality of the estimated erosion, the model's output, based on these input data is very relevant for improving the reliability of decision-making based on erosion estimates.

During simulation, all input data are combined throughout the model and influence its output in different ways. Therefore, the DQI of each input data affects the DQI of the model's output according to the associated sensitivity.

Thus, it is demanded to know the sensitivity of each model's input data to evaluate the DQI of its output. It is performed by sensitivity analysis, which assesses the impact of each variable on the outcome of the erosion model. Variables with high sensitivity have a significant effect on the model's estimates, while variables with low sensitivity have a lesser impact.

Sensitivity can be estimated using advanced methodologies such as Shapley or GSA, which quantify the contribution of each variable to the model's uncertainty regarding model's nonlinearities and interactions. This analysis is essential for prioritizing which variables require closer assessment and monitoring, particularly when complemented by qualitative evaluation. [19][20][21].

Using Shapley, the sensitivity parameters (SP) presented in table 3 are obtained. As shown, the oil rate and the pressure are the most relevant input data, followed by the temperature, the sand content, the gas-oil ratio and the BSW. Other variables are not relevant.

Table 3. Sensitivity parameters

Input data	SP
Oil rate	28.97
Pressure	28.01
Temperature	13.04
GOR	4.24
BSW	0.97
Oil density	0.00
Oil viscosity	0.00
Gas viscosity	0.00
Sand content	6.25

By combining the sensitivity parameter with the DQI of each model's input data according to equation 1, it is possible to evaluate a DQI of the model's output.

$$DQI_M = \frac{\sum_{i=1}^{N} DQI_i \cdot SP_i}{\sum_{i=1}^{N} SP_i} \quad (1)$$

where:

- DQI_M: data quality indicator of model's output
- DQI_i: data quality indicator for each model's input data "i"
- N: number of model's input data
- SP_i: sensitivity parameter of the input parameter "i".

Thus, evaluating a DQI of each model's input data and a SP for model parameter, equation 1 allows the evaluation of data quality of the model output, as shown in table 4.

Table 4. Data Quality of DNV-O501 model output

Erosion estimate:	3.28 mm
Data quality indicator	0.77
Data quality class:	Intermediate quality data

The advantages of evaluating the model´s output DQI are associated with providing information about the reliability of the model's estimate. It is a measure of how trustworthy the model output is, as well as the decision-making based on it.

Moreover, it can be used to monitor model reliability over time, and to provide insights about how to improve model estimates assertiveness by means of acting on bad quality data, or its source.

VII. Conclusion

This work explores the application of FAIR principles and data quality assessment within the context of digital twins. By utilizing an unique database where all data is conditioned and ready for use by each model, it becomes possible to assess the quality of data inputs for each model.

Additionally, with the help of sensitivity parameters evaluated through methods like Shapley, the quality of the model's outputs can also be assessed. The ability to evaluate data quality indicators for both inputs and outputs enhances the amount of reliable information, improving the accuracy and trustworthiness of the model's estimates and predictions.

Ultimately, this leads to greater reliability, higher assertiveness and reduced uncertainty in decision-making, making maintenance planning more reliable and effective.

References

[1] G. Pauli, M. B. Demay, A. M. da Mata, S. S. Rodrigues, J. M. Xavier, J. de O. Braga, G. D. Donatelli, E. Margotti. "On the Impact of Temporal Resolution on Nonlinear Model Accuracy for Predicting Wear Due to Solid Particle Erosion in Digital Twins of Oil and Gas Equipment." Paper presented at the Offshore Technology Conference Brasil 2023 (OTC Brazil 2023). https://doi.org/10.4043/32797-MS

[2] S. S. Rodrigues, J. de O. Braga, J. M. Xavier, G. Pauli, A. M. da Mata, E. Margotti, M. B. Demay, G. D. Donatelli. "Digital Twin of Subsea Assets as a Tool for Integrity Management and Risk-Based Inspection: Challenges and Perspectives." Paper presented at the Offshore Technology Conference Brasil 2023 (OTC Brazil 2023). https://doi.org/10.4043/32757-MS

[3] J. D. Hochhalter, W. P. Leser, J. A. Newman, E. H. Glaessgen, V. K. Gupta, V. Yamakov, S. R. Cornell, S.A. Willard, G. Heber. "Coupling Damage-Sensing Particles to the Digital Twin Concept." NASA/TM–2014-218257. 2014.

[4] F. Tao, Q. Qi, L. Wang, A.Y.C. Nee. "Digital Twins and Cyber–Physical Systems toward Smart Manufacturing and Industry 4.0: Correlation and Comparison", Engineering, v. 5, 2019, p-p 653-661, https://doi.org/10.1016/j.eng.2019.01.014.

[5] Q. Qi., F. Tao. "Digital Twin and Big Data towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison". IEEE Access 6 (2018): 3585–93. https://doi.org/10.1109/access.2018.2793265.

[6] E. Glaessgen, D. Stargel, "The Digital Twin Paradigm for Future NASA and Air U.S. Force Vehicles." 2012. https://doi.org/10.2514/6.2012-1818.

[7] Tao, F. 2018. "Digital twin-driven product design, manufacturing and service with big data". Int. J Adv Manuf. Technol. 94: 3563–3576. https://doi.org/10.1007/s00170-017-0233-1.

[8] DNV. "DNVGL-RP-O501: Managing sand production and erosion". 2015.

[9] J. Top, S. Janssen, H. Boogaard, R. Knapen, G. Simsek-Senel, "Cultivating FAIR principles for agri-food data". Comput. Electron. Agric. 196. 2022. https://doi.org/ 10.1016/j.compag.2022.106909.

[10] N. N. K. Krisnawijaya, B. Tekinerdogan, C. Catal, R.van der Tol, Y. Herdiyeni, "Implementing FAIR principles in data management systems: A multi-case study in precision farming", Computers and Electronics in Agriculture, v. 230, 2025, https://doi.org/10.1016/j.compag.2024.109855.

[11] Gofair. "FAIR Principles". https://www.go-fair.org/fair-principles/. 2025.

[12] A. N. de la Hidalga, J. Goodall, C. Anyika, B. Matthews, C. Richard A. Catlow. "Designing a data infrastructure for catalysis science aligned to FAIR data principles". Catalysis Communications, v. 162, 2022, https://doi.org/10.1016/j.catcom.2021.106384.

[13] DNV. "DNV-RP-0497: Assurance of data quality management". 2023.

[14] DNV. "DNV-A204: Assurance of digital twins". 2023.

[15] ISO. "ISO 8000-8: Information and data quality: Concepts and measuring". 2015.

[16] M. B. Demay, G. D. Donatelli, A. L. M. de Oliveira, S. S. Rodrigues, "On the influence of uncertainty in risk analysis based on digital models", Measurement: Sensors, https://doi.org/10.1016/j.measen.2024.101510.

[17] D. Curto, F. Acebes, J.M. Gonzales-Varona, D. Poza, "Impact of aleatoric, stochastic and epistemic uncertainties on project cost contingency reserves". 2022. Int. J. Production Economics 253.

[18] JCGM, "JCGM 100: evaluation of measurement data — guide to the expression of uncertainty in measurement (ISO-GUM)", 2008.

[19] M. B. Demay, G. D. Donatelli, A. L. M. de Oliveira, P. H. Z. Machado, S. S. Rodrigues. "On the meaning of sensitivity analysis". Measurement: Sensors, 2024, https://doi.org/10.1016/j.measen.2024.101511.

[20] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto. "Sensitivity analysis in practice: a guide to assessing scientific models", v. 1.. New York: Wiley. 2004.

[21] H. M. Wainwright, S. Finsterle, Y. Jung, Q. Zhou, J. T. Birkholzer, "Making sense of global sensitivity analyses", Computers & Geosciences, v.65, 84-94, 2014, https://doi.org/10.1016/j.cageo.2013.06.00