ICA2025OSLO
14th IWA Specialised Conference on
Instrumentation, Control and Automation

# Metadata: challenges and opportunities to support decision making in the wastewater sector

J. Alferes*, D. Aguado**, M.V. Ruano***, O. Samuelsson****, Q. Plana***** and K. Villez******

*VITO – WaterClimateHub, Ostende, Belgium, janelcy.alferescastano@vito.be
**IIAMA, Universitat Politècnica de València, València, Spain, daaggar@hma.upv.es
***Chemical Engineering Department, Universitat de València, Burjassot, Spain, m.victoria.ruano@uv.es
****IVL Swedish Environmental Research Institute, Stockholm, Sweden, oscar.samuelsson@ivl.se
*****Eurecat – Technology Center of Catalonia, Barcelona, Spain, queralt.plana@eurecat.org
******Oak Ridge National Laboratory, Oak Ridge, TN, USA, villezk@ornl.gov

**Abstract:** Metadata describing sensor signals is critical to foster confidence in the data generated and in any further use of the data, thus unblocking its value. This becomes especially relevant within the digitalization process that the wastewater sector is going through, with huge amount of data being generated. Nevertheless, as of today, this is limited by a lack of frameworks describing: what metadata is, how metadata should be organized, common practices to define metadata describing sensor conditions, and a comprehensive overview of data analytical methods that can describe sensor signals and the associated metadata that can be generated. This paper contributes to bridge the current gap towards best practices associated with metadata and data generation for online sensors, in the context of water resource recovery facilities (WRRFs). A complete overview of metadata to describe signal generation, sensor condition and signal quality is provided. Special emphasis is given to metadata providing information on the quality of the data collected. Current practices for selecting and implementing data quality assessment tools rely on inefficient trial-and-error approaches. To support a more systematic decision making towards an optimum method selection, a practical framework for selection and application of data quality assessment methods is also presented. The value of metadata describing sensor signals lies in making data useful for a particular purpose within the organization. Although important challenges remain, metadata is strategic as part of any digital roadmap. Main steps towards such metadata transformation are also proposed to turn data into valuable information to better operate WRRFs.

**Keywords**: data-quality assessment; digitalization; metadata; online sensors

Digital transformation is and will be playing a crucial role in the transition journey for the wastewater sector, towards more efficient, robust and sustainable systems. New monitoring technology generating rich datasets, advances in ICT and automation, real-time modelling, digital twins and data-driven approaches, offer huge potential to develop customised solutions and to exploit the value of data (Grievson et al., 2022; Bettin A., 2023). However, with the opportunities also come the challenges. Currently, standard best practices for data collection, data storage and data management are lacking, with mostly silo systems, making particularly difficult data exchange and interoperability. This also leads to data graveyards and the lack of 'good' and useful data to be properly used within data driven solutions, real-time control, process optimization and decision making (Therrien et al., 2020).

Although most water resource recovery facilities (WRRF) already collect large data volumes today, systems, best-practices and knowledge to tactically handle these data are often limited or even non-existent. One of the reasons is that a systematic description of sensor signals, including sensor signal quality assessment, is often absent. Such information about the data or 'metadata' is key to convert meaningless streams of raw data into fit for purpose data. Metadata is also invaluable when searching through volumes of historical data, trying to identify certain process conditions. Additionally, harsh measurement environments make online sensors prone to faults, affecting the reliability of the data collected and therefore its proper use (Samuelsson, 2017). This is often underestimated. Metadata, including a quantification of the data quality becomes crucial and an enabler for a proper management and exploitation of the whole data chain (Figure 1). This paper puts forward best practices to understand, collect, organize and implement data and metadata within WRRFs, based on existing standards and state-of the-art algorithmic tools, with focus on practical applications, aiming to facilitate improved decision-making by leveraging sensors data available in WRRFs, thereby contributing to their digital transformation process.

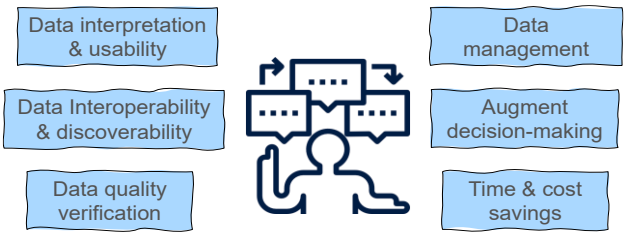
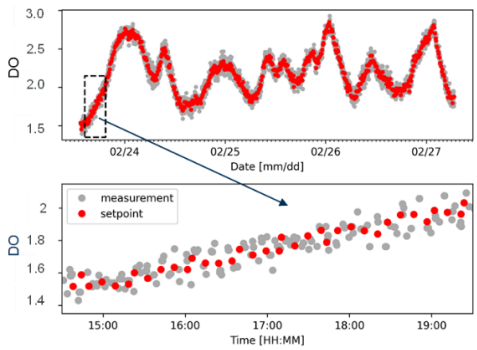



**Figure 1** Metadata turning raw data into representative, reliable, interpretable & useful data

Metadata provides information about the data and allows turning large volumes of raw data into representative, reliable, interpretable and useful data. Figure 2 illustrates an example of metadata giving information on the sensor signal generation for a dissolved oxygen (DO) measurement.




| | |
|---|---|
| Value | 2.0 |
| Unit | mg/L |
| Time stamp | 2020-02-23 18:50:17 |
| Measurement principle | Luminescence-based sensor |
| Location | Aeration basin 3 |
| Sensor model | Hach LDO |
| Accuracy (RSME) | 0.2 |
| Purpose | Control loop |

**Figure 2** Example of metadata for a DO measurement

At wastewater utilities, data produced is obtained through a variety of devices, generating highly variable and heterogeneous data in its structure. Dealing with this divergency of data formats can be a challenge when storing or interpreting the data. A comprehensive overview of metadata types is proposed by grouping metadata into three main categories: (1) metadata describing signal generation, (2) metadata describing signal quality, and (3) contextual information in the form of annotations. Figure 3 summarizes the main categories. A priority level has been defined (examples illustrated in Figure 4): Level 1 (basic) – a signal cannot be understood without this information, Level 2 (advanced) – useful for sensor validation and within automated tools for data quality assurance and control and, Level 3 (Future proof) – useful for future use of the collected data.

1
I. Signal production
II. Signal producing device
III. Signal modification

2
I. Measurement errors, through observed symptoms
II. Measurement errors, through their root causes
III. Diagnosis

3
Metadata providing **contextual** information relevant for later use

**Figure 3** Proposed metadata types

| METADATA | Examples | | Review frequency | Priority level |
|---|---|---|---|---|
| | DO sensor | pH sensor | | |
| Timestamp origin | Analog-digital converter (PLC integrated) | SCADA system | Years | 1 |
| Temporal resolution | 15s | 1s | Years | 2 |
| Numeric resolution | 0.3 x $10^{-3}$ (16-bit) | Not available | Years | 3 |
| Sensor location | Aeration basin 3 in lane 2 | Primary clarifier outlet | Months -Year | 1 |
| Measurement unit | $g/m^3$ | Not applicable (dimensionless) | Years | 1 |
| Measurement range | 0-20 | 2-12 | Years | 1 |
| Variable range | 2-6 | 5-7 | Months-Year | 2 |
| Transmission mode | Digital: Modbus | Analog: 4-20 mA | Years | 2 |
| Measurement principle | Luminescence-based sensor | Ion-selective electrode | Years | 3 |
| Purpose of measurement | Automatic feedback control of DO | Warning for industrial influent | Months - Year | 3 |

**Figure 4** Examples of metadata describing signal production devices

In view of a proper use of the data, special attention must be paid to the quality of the data being collected, meaning that data is assessed, measurement errors identified and when needed, quantified and corrected. Efficient methods to ensure and validate the quality of the data are then essential. Figure 5 summarizes the metadata defined to describe sensor data quality. With the goal of better use and take advantage of available data quality assessment techniques (e.g. Newhart et al., 2019; Teh et al., 2020), a clear distinction is made between identified causes for a fault or abnormal data and the symptoms that conducted to its detection.

| Metadata | Description | | | |
|---|---|---|---|---|
| (1) Measurement errors through their observed symptoms | Errors as deviations from reference values | Static – basics: accuracy, trueness & precision | Static – advanced: repeatability & reproducibility | Dynamics: response time, shocks, drift, outliers |
| (2) Measurement errors through their root causes | Systematic errors: calibration error, fouling, instrument error… | Gross error: wrongful reading, recording, computational | Random errors | |
| (3) Diagnosis | Experimental methods and algorithm tools to detect and help identify the root cause of sensor data quality degradation | | | |

**Figure 5** Metadata describing sensor data quality

Measurement errors have several causes, including hardware malfunctions, human error (e.g. wrong calibration), measurement environment (e.g. formation of fouling and scaling), and changes in the measured conditions. Unfortunately, matches between root causes and affected metrics or observed symptoms are not always evident. Classical measures to quantify the symptoms are often insufficient to isolate a unique root cause as many symptoms are not unique to a single root cause, so multiple potential root causes must be investigated. Both experimental methods and algorithmic tools are then proposed to eliminate root causes and refine the diagnosis. Concerning algorithmic tools with potential use for WRRFs applications, two main approaches are presented: (1) methods rooted in mechanistic understanding and (2) methods rooted in statistics and data analysis. In the first case, mechanistic knowledge of the process is known and therefore can be used (e.g. sanity checks, mass balances). In the second case, information from historical data can be exploited by implementing methods that are driven by the data (e.g. supervised and unsupervised approaches). From the practical point of view, the choice will be mostly conditioned by (1) the available data and therefore the type of learning that is possible (e.g. availability of labelled data sets), and (2) the desired behaviour of the data validation implementation. To facilitate the decision-making process for selection and implementation, a framework for the application of data quality assessment tools is presented in Figure 6.

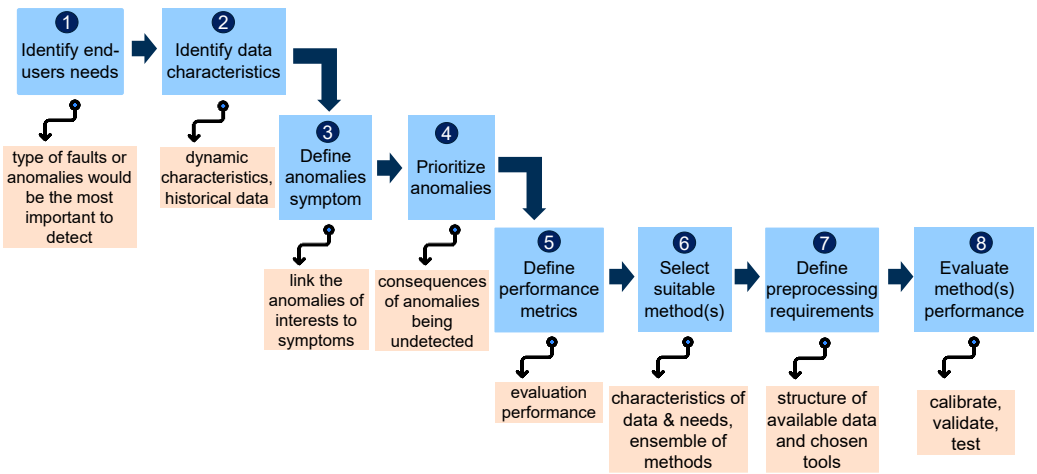



**Figure 6** Practical guidelines for selection of data quality assessment methods

As a good practice, domain experts and data end-users should always be involved in setting-up the data assessment framework and in evaluating its performance. Regardless of the chosen approach, it is important to take into account that all methods have their own strengths and limitations. An efficient, reliable, and meaningful evaluation may require to combine several approaches. Some examples are the use of hybrid methods (integrating data-driven and mechanistic models), and the combination of univariate and multivariate methods (detecting issues in individual sensors and situations affecting the whole system). Even if a significant number of methods are available, there is still an important gap in bringing these methods in practice, with a lack of comprehensive and systematic methodology (Russo et al., 2020). In reality, data quality assessment and any data processing included in a data pipeline are case and purpose specific.

Careful data management practices will include the collection of metadata to know with confidence where, when, how, and why sensor signals were produced, assess the data quality produced by a sensor and adequately frame sensor signals in the context in which they were produced. Thus, metadata becomes a critical part of the bigger picture on the digital transformation of any water utility. This transformation should follow a two-step procedure as proposed in Figure 7, touching technical but also organizational aspects. Upcoming efforts should be focused on creating accurate, consistent, linked, user-centred, and high-quality metadata, facilitating interoperability across units, utilities, and stakeholders.

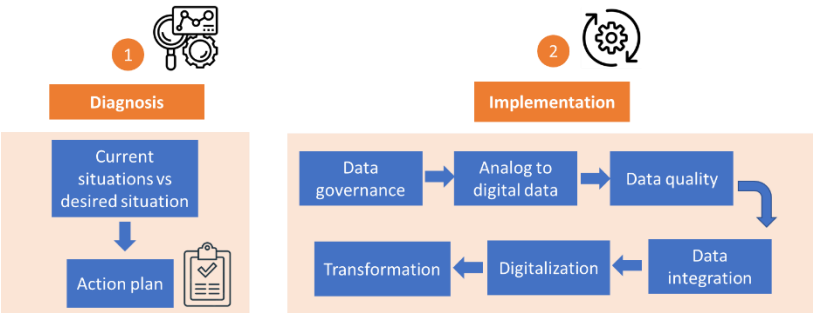



**Figure 7** Main steps towards a data and metadata transformation

**Acknowledgement**
The authors acknowledge the IWA association for supporting the MetaCO task group and all contributors to the STR 'Metadata Collection and Organization in Wastewater Treatment and Wastewater Resource Recovery Systems' on which this paper is based.

## REFERENCES

Bettin A. (2023) Digital transformation for the water industry: how a data-driven business intelligence platform can improve operations. Water Practice and Technology 18 (7), 1599–1607. doi: https://doi.org/10.2166/wpt.2023.091

Grievson O., Holloway T. and Johnson B. (2022) A Strategic Digital Transformation for the Water Industry. IWA Publishing, 2022. https://doi.org/10.2166/9781789063400

Newhart K.B., Holloway R.W., Hering A.S. and Cath T.Y. (2019) Data-driven performance analyses of wastewater treatment plants: A review. Water Research, 157, 498-513. https://doi.org/10.1016/j.watres.2019.03.030.

Russo S., Moritz L., Hao W., Matthews B. and Villez K. 2020 Active learning for anomaly detection in environmental data. Environmental Modelling and Software 134.

Samuelsson Oscar. 2017. PhD thesis on Sensor Fault Detection and Process Monitoring in Water Resource Recovery Facilities. Uppsala University.

Teh H.Y., Kempa-Liehr A.W. & Wang K.IK. (2020) Sensor data quality: a systematic review. J Big Data 7, 11. https://doi.org/10.1186/s40537-020-0285-1

Therrien J.D, Nicolaï N, Vanrolleghem P.A. (2020) A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. Water Sci Technology , 82 (12), 2613–2634. doi: https://doi.org/10.2166/wst.2020.393.