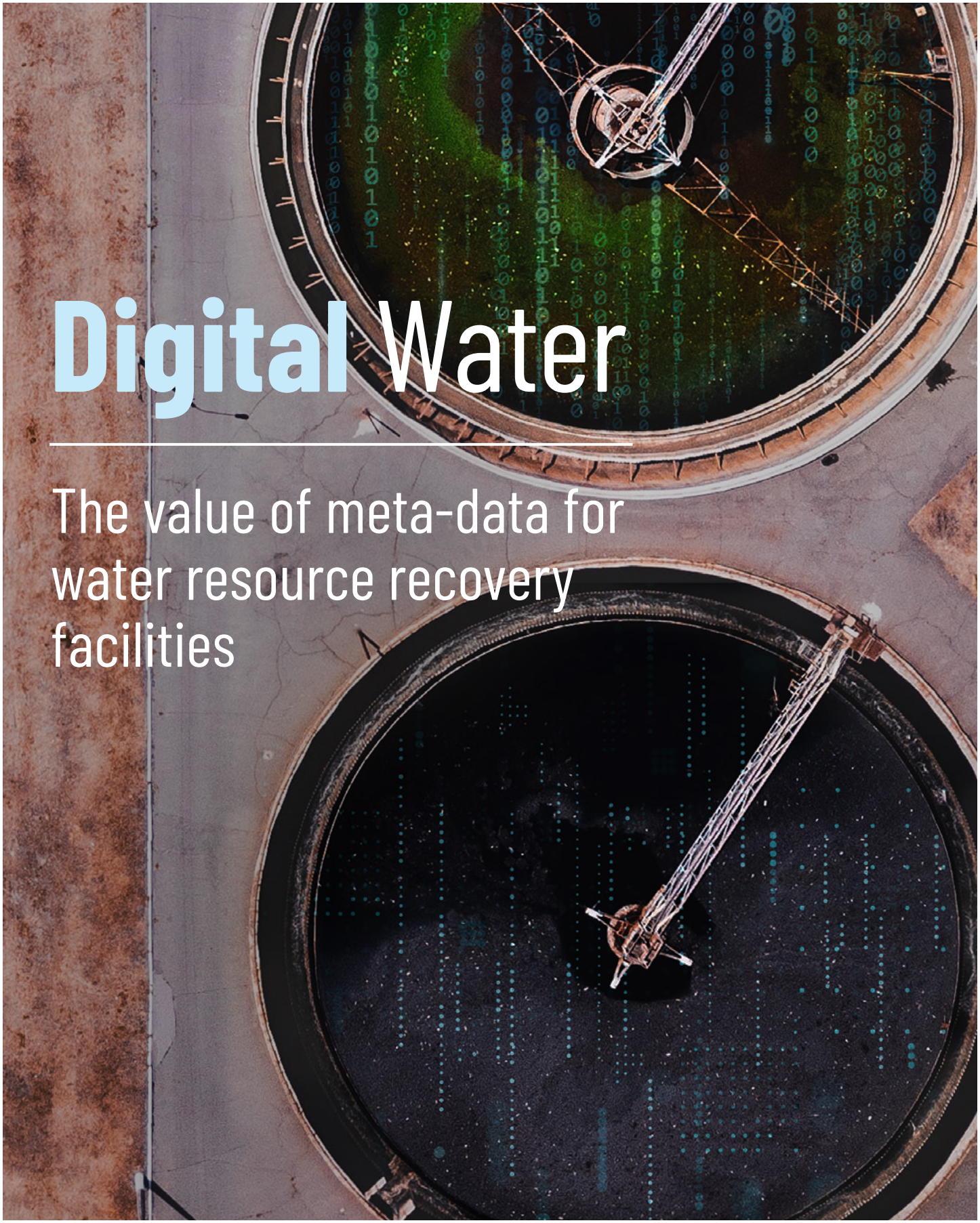
the international water association



Digital Water

The value of meta-data for water resource recovery facilities

the international water association

Digital Water

The value of meta-data for water resource recovery facilities

Authors

Daniel Aguado *Associate Professor, Institut Universitari d'Investigació d'Enginyeria de l'Aigua i Medi Ambient (IIAMA), Universitat Politècnica de València, València, Spain*

Frank Blumensaat *Senior Research Fellow, ETH Zurich/Eawag, Institute of Environmental Engineering, Chair of Urban Water Management Systems, Zurich, Switzerland*

Juan Antonio Baeza *Full Professor, Department of Chemical, Biological and Environmental Engineering, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain*

Kris Villez *Sr. R&D Staff Member, Oak Ridge National Laboratory, Oak Ridge, TN, USA*

María Victoria Ruano *Associate Professor, Chemical Engineering Department, Universitat de València, Burjassot, Spain*

Oscar Samuelsson *Researcher, IVL Swedish Environmental Research Institute, Sweden*

Queralt Plana *R&D Engineer, modelEAU, Université Laval, Québec, Canada*

Foreword

With the water sector having joined the digital revolution, water utilities and companies now need to incorporate digitalisation into their vision for the future of their infrastructure and operations. The journey towards a smarter future will not be without challenges, but it will lead to an increase in performance, and, in the longer run, more sustainable and inclusive water management.

Our sector is on the front line against many current and future challenges facing the world (such as climate change, population growth, water scarcity, etc). Digitalisation can help to address these challenges by optimising operations, improving performance, and reducing uncertainty. However, to fully embrace digitalisation, it will be crucial to understand how to maintain digital tools and how to interpret and manage the obtained data.

More specifically, one important challenge that needs to be addressed is the management of a conspicuous amount of data obtained using digital tools. It is crucial to identify data which are useful for tackling existing problems and to separate them from data that will be needed to address future challenges.

This latest contribution from the IWA digital water white paper series discusses the importance of meta-data and how we ensure we have access to data mines, rather than data graveyards. Doing so will enable a possible route to address the meta-data challenge- i.e. understanding which data will be useful to address future, yet unknown, challenges.

The IWA is leading the transformation towards a smarter global approach to water management. Through the IWA Digital Water Programme, the association provides a platform where water professionals can exchange knowledge, discuss challenges and develop effective solutions. This includes producing this digital water white paper series to provide insights into how our sector can take advantages of technological developments, enabling it to adapt and be more robust in the context of the pressures of global change.

At IWA, we believe it is essential to generate and share knowledge and best practice, in order to develop effective solutions that can ensure the ongoing improvement of our sector. Doing so will improve productivity, reduce our carbon footprint, and ultimately lead to a more effective and inclusive response to future challenges.

Kalanithy Vairavamoorthy
Executive Director of the International Water Association

Summary

Meta-data refers to descriptive information essential to convert large volumes of raw data into useful resources. With the advance of digitalisation in the water sector, it is fundamental to avoid data graveyards and, on the other hand, using collected data to address current and future problems. This white paper focuses on the crucial role that meta-data has in responding to future and possibly unpredictable challenges. The aim of this document is to present the 'meta-data challenge' and to highlight the need to consider meta-data when collecting information as part of good digitalisation practices.

Introduction

As water resource recovery facilities (WRRFs) enter the era of big data, they are naturally confronted with the challenges of integrating smart actuators, sensors, and autonomous control systems in a sensible and transparent manner. One aspect that remains an important burden to bear by water utilities is the storage and management of sensor data in view of later use. To enable data interpretation beyond the original time of data collection, it is crucial that the collected sensor data is augmented with an adequate description, i.e. meta-data. Indeed, data collected today are expected to be useful in the future to respond to even more complex operational challenges and new demands for the environmental impacts, the produced effluent quality, and resource efficiency. Given that those future challenges are unknown, it is particularly challenging to define the required meta-data for a generation of future-proof data mines, as opposed to data graveyards, and ensure its collection in a timely manner. In this white paper, we highlight the most important aspects of this meta-data challenge and we provide arguments and early solutions leading towards harmonised data collection and interpretation. This white paper represents the first of many outcomes of the IWA Task Group on Meta-Data Collection and Organisation (MetaCO TG) which has been supported by the International Water Association since 2020. The MetaCO TG is currently working on a Scientific and Technical Report (STR) on meta-data which will complement the information presented in this report (scheduled for 2023). Together with this white paper, the STR will provide more information on the meta-data challenge.

The need for meta-data

In the last decades, the water sector has undergone an instrumentation revolution. For example, the measuring of dissolved oxygen was introduced first to WRRFs to augment the existing capabilities of flow and level sensors in the 1980s. Since then, the diversity of available sensors has steadily increased. The challenges and opportunities in collecting 'big data' are often categorised into the following four 4 V's: Velocity, Volume, Variety, and Veracity. Thanks to increasingly efficient communication techniques and extreme reductions in data storage costs, data collection has become extremely scalable. This means that today's WRRFs are now mastering the first two of the four V's, Velocity and Volume [1].

Recent developments in data mining, machine learning, and optimisation, enabled by virtually endless computational power and algorithms for computer-based learning, have been met with enthusiasm in the water sector. Many are enticed by new capabilities of computer-aided decision-making both at an operational and managerial level. However, many attempts in advancing automation from the increasingly large data streams invite a hard confrontation with the other two V's of big data: Variety and Veracity [2-3]. Human intelligence and smart routines are still needed to categorise, structure, homogenise, and convert data into valuable information. Indeed, this important step easily demands 40% of the costs in most consultancy and data science projects, both in the wastewater treatment sector and others [4-7]. This cost is largely associated with the need to triage the available data (i.e. separate garbage data from data fit for purpose) to avoid the commonplace Garbage-In Garbage-Out problem, which is now more obvious than ever.

The authors of this report believe the cost of this task can be reduced drastically if routine data collection and management practices are updated to support data-intensive decision-making and automation. More specifically, existing data should be augmented by providing information on the original purpose, the data-generating devices, the quality, and the context of these data. This kind of descriptive information is known as meta-data and is an essential ingredient to turn large volumes of raw data into actionable information. Indeed, detailed knowledge about the measurements are needed for sound and creative data analysis, so as to guarantee an impact on design and operational decisions [6]. Unfortunately, there are no wastewater-specific guidelines available to the production, selection, prioritisation, and management of meta-data.

Garbage-In Garbage-Out

Why data-analytical tools are picky

There are several reasons why simply collecting more and more data is not sufficient. This holds both when mechanistic models (e.g., physics-based) or empirical models (e.g., including machine learning and other data-based models) are used for data interpretation. Here are a few reasons and required adjustments to resolve them:

1. Typical data include measurements of suspect quality. It is common to observe the symptoms of short-lived sensor faults in the form of outliers, spikes, high-noise features, and sustained deviations from the reference measurements. Moreover, water quality sensor signals are prone to sustained faults, often due to calibration errors and drift. Including all data without elimination or correction of low-quality measurements will lead to faulty models. A key requirement is that the data used for model identification should be of high quality, either through proper management of the data collection system or through a well-established data refinement process with offline or online data validation and reconciliation tools.

2. Often, the data available for model identification does not correspond to the conditions under which the model will be deployed. Consider the flow and influent composition data from a WRRF under dry and wet weather. These data should be clearly separated to obtain a reliable model for the typical influent of a WRRF under these distinct operational conditions. Therefore, another requirement is that the data used for model identification should be representative for the deployment conditions.

3. While data might be voluminous, the patterns one wants to analyse are often rare (e.g., toxic spills, rain events). Indeed, one of the many promises of machine learning is that it can help detecting and diagnosing rare events. To make this simpler, a sufficiently large data record corresponding to the events of interest should be available or any imbalance in the frequency of events should be accounted for through the provision of detailed system knowledge.

Today, the requirements mentioned above lead to a necessary but cumbersome triage before data analysis can take off. Human experts crawl through the large volumes of data and modify, select, and annotate the data to enable a correct execution of the computer-aided prediction or optimisation task. This is a tedious effort and often includes subjective assessments by a domain expert. Meta-data, if available, can help. First, informative meta-data can assist by automating the triage to a high degree. Second, structured meta-data reduces the need for subjective assessment, in turn increasing trust in algorithmic predictions and decision-making. Since any measurement-based algorithm relies on representative, reliable, and interpretable data, data sets should be judged by virtue of the provided meta-data, next to the conventional measures of sensor signal quality, such as trueness, precision, and response time (see [8] for definitions). Figure 1 illustrates how measurement values alone are not sufficient to reap the benefits of intensive data collection systems.

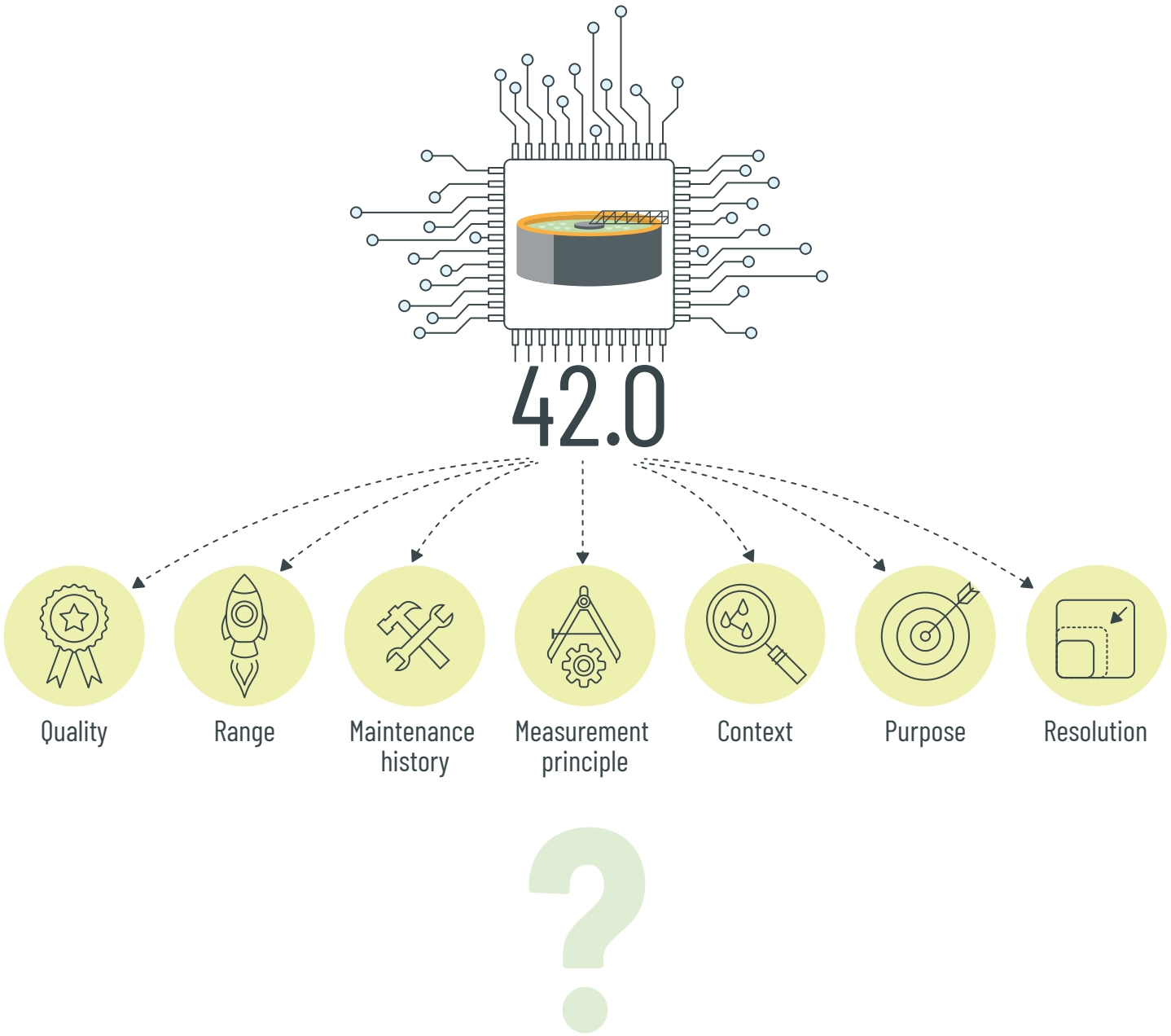


Figure 1. Can we interpret the provided measurement equal to 42.0? We need to know a lot more to evaluate the information contained in sensor signals. The kind of descriptive data we need for interpretation is known as "meta-data" and includes the purpose of measurement, the measurement principle, the temporal and measurement resolution, sensor maintenance history, indicators of signal quality, and the spatial and temporal context of the measurement.

For data triage purposes, a good data set will include meta-data, about the following aspects:

1. Data-generating system – Information describing every step of the data collection process, including information about (a) the purpose of data collection, (b) the sensor hardware (e.g., measurement and temporal resolution, measurement unit, measurement principle, manufacturer, sensor model etc.), (c) signal management, including recording, transmission, and storage, and (d) data refinement, including all modifications of the data after data collection. This kind of meta-data enables to select data that are fit-for-purpose and is often already available from the sensor devices themselves through a modern digital communication system (e.g., Ethernet).

2. Data quality – Detailed information about the sensor signal quality and is based on (a) a digital record of all sensor calibration, validation, and verification events for every sensor, (b) description of individual records that are suspect (e.g., outliers, spikes), and (c) descriptions of sustained periods with poor data quality (e.g., due to calibration errors, lack of maintenance, drift, and other malfunctions). Such descriptions can be obtained through manual data annotation by a domain expert but, importantly, also by careful deployment of data-analytical tools, in turn leading to quantitative data quality assessment and quality control. This means that also (d) the output from the methods used to collect this information (e.g., algorithmic analysis, standard operation protocols, expert annotation etc.) should be included as meta-data. This kind of meta-data enables to identify data that satisfies the required data quality.

3. Contextual information – Information describing the circumstances outside of the plant that may influence the interpretation of signals recorded on the plant. This could be information on process mode, local weather, including seasonal changes or storm weather, or meaningful changes in the structure and operation of upstream infrastructures (e.g., sewer). It also includes information on rare events, including operating failures, social gatherings, or toxic spills. Quite often, this kind of information is provided by technical and operational staff. This type of meta-data allows to pick data that is relevant to the task at hand, i.e. that is informative and relevant.

Unfortunately, the meta-data described above are rarely available. For example, descriptions of the procedures (e.g., sensor maintenance) might be missing, results from sensor validation may not be logged, and the circumstances under which data were collected (e.g., storm weather) could be unknown. This kind of information is indispensable however for data-intensive prediction and optimisation of the performance of WRRFs. Without it, one critically relies on the memory of on-site staff to

interpret the available data. Within a year, historical data can become useless for most data-based tasks as personal memory fades and the collected data meet their expiry date. This loss of valuable descriptive information can quickly turn information-rich measurements into a data graveyard. Furthermore, this lack of descriptive information often only becomes apparent a long time after the original measurements were collected. Thus, effective data governance not only requires that one can answer today's questions based on data but also to manage the collected data in such a way that future yet unknown questions will be answered reliably too. To account for these unknowns, as well as for an ageing work force, to empower staff members, and to assure the long-term utility of historical data records (e.g., decades) for important decisions at operational and managerial level, collection of meta-data of the types discussed above should become a routine matter.

Structuring meta-data

Even when meta-data is safeguarded for later use, it may be challenging to wield it. Indeed, meta-data frequently resides in a vast array of design specifications, manuals, protocols, and spreadsheets, often stored in separately managed databases and folders (i.e., silos). The location of, and access to these data is often managed in an ad hoc manner. To enable the envisioned triage of data with a highly automated process, meta-data should be accessible and stored in a way which allows for an easy navigation. To achieve this, meta-data needs to be stored in a structured manner and routinely kept up to date. Where feasible, a centralised system for meta-data storage will be helpful to manage access while ensuring accuracy and completeness.

The definition and integration of meta-data is often a hurdle and typically not part of off-the-shelf software products used in the water sector. For this reason, it is important to focus on ways this can be achieved. These include the generation of the identification of a primary data source, which points to the most accurate and complete version of all data and meta-data. This primary data acts as a single-source-of-truth and engenders a shared and unambiguous understanding of the most current data and information available to all data users. Naturally, identifying a primary data source implies a well-calibrated appreciation of the need for good governance of data, information, models, and software. In turn, this means that these changes affect almost everyone in the organisation, thus requiring a careful alignment of objectives and needs as part of good digitalisation practice.

Sensor maintenance revisited

Leveraging existing quality assessment procedures for quick wins

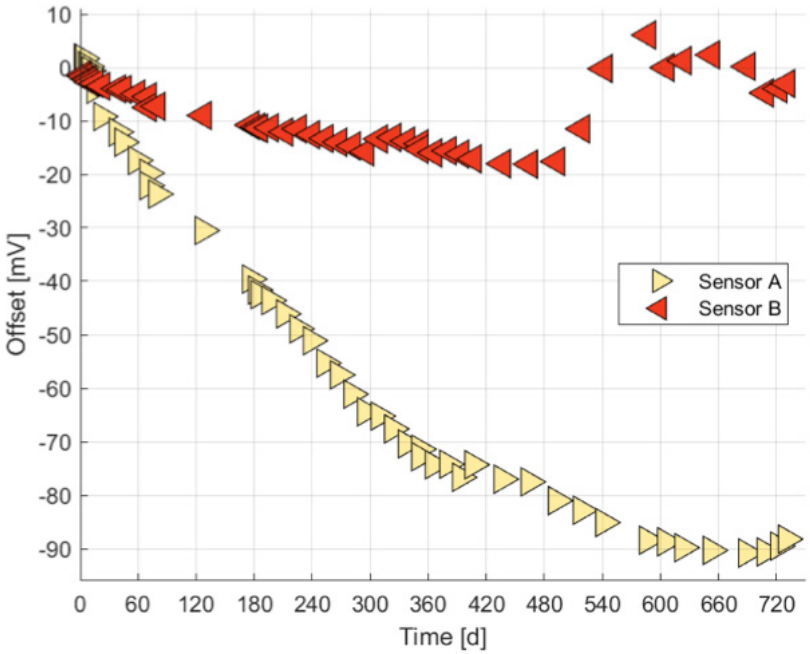


Figure 2. Effects of wear-and-tear on two pH sensors. The sensor offset (electrode potential at pH=7) diminishes gradually as time progresses for sensor A. This is attributed to drift of the reference electrode and can be accounted for by calibration. The accumulated drift is about -90 mV by the end of a two-year period, which amounts to a shift of roughly 1.5 pH units in absence of calibration. In sensor B, the offset increases after 500 days of use. This is attributed to irreversible damage to the sensor and requires sensor replacement. Produced with data from [9].

Many utilities have protocols in place to check the validity of sensor signals on a regular basis. More often than not, a reference measurement is used to determine whether a maintenance action, such as intensive cleaning or calibration, is needed. However, this kind of reference measurement is rarely recorded. To illustrate how a modest enhancement of existing quality assessment and control practices can lead to useful meta-data, we take two series of reference measurements from [9].In Figure 2, one can see the offset of two pH sensors as a function of time during their 2-year deployment. This offset is the measured potential of the sensor in a calibration solution with pH 7 and is available as part of the sensor calibration curve stored within a typical transducer. One sensor (sensor A) exhibits a monotone, decreasing profile. This decrease is easily explained by drift of the reference electrode. This drift is compensated in practice by regular calibration. In contrast, the second sensor (sensor B) exhibits an increase of the offset after 500 days. This is due to irreversible wear-and-tear of the sensor and cannot be corrected with calibration. Importantly, making a distinction between the offset around 180 days (-11.4mV), which is explained by normal drift, and the similar offset value at 510 days (-11.5mV), explained as a result of damage, is only feasible thanks to the whole history of offset values. Without these meta-data, such a diagnosis is infeasible.

We therefore encourage the systematic recording of this kind of meta-data before and after every maintenance action already executed on the plant, including cleaning, calibration, and part replacements. This will enable an accurate and timely response to sensor wear-and-tear, thus improving data quality. In the long run, it can also produce valuable information to implement preventive sensor maintenance and decide on the best sensing hardware, particularly by quantifying the trade-offs between the cost of sensor hardware against the obtained data quality and costs of maintenance actions.

The meta-data work force

As mentioned above, effective digitalisation requires cultivation of good meta-data management practices, many of which can be automated through careful selection and structured management of meta-data. This has produced a range of novel roles in the wastewater sector, with names like data steward, data engineer, or chief data officer, highlighting the need for in-house expertise in data handling, managing of expectations regarding digital transformation, and translation of opaque computational concepts into a common sense language. These experts can be extremely helpful to turn stale databases into an effective source of information for operations and investment decisions. They should be tightly integrated into the existing work force, to facilitate early adoption of evidence-based decision-making and to ensure alignment of expectations and objectives across the organisation. In the future, we hope specialised teams and staff can reach across the following topics of relevance, which are often handled by different subject matter experts today, each with their own isolated terminology:

1. Add and integrate new devices into an existing control system.
2. Manage and optimize data collection systems in multi-purpose settings, e.g., data for real-time control, for reporting, for model construction and validation and for planning major upgrades.
3. Manage and optimise sensor data quality aided with basic and advanced data validation tools.
4. Augment existing data streams with meta-data, as described in a highly automated fashion.
5. Merge the meta-data needs for WRRF operations, algorithmic requirements, and current sensor data collection routines.

Take home message

Getting started with meta-data

Following enthusiastic responses to breakthroughs in artificial intelligence, robotics, and machine learning, it is increasingly clear that reaping the benefits of intensive data collection requires augmentation of sensor signals with descriptive information. The need for this descriptive information, called meta-data, and the challenges to obtain and manage it underscore that there is no free lunch. Effective data governance includes the provision of high-quality meta-data and will make the difference between failures and successes in data-intensive system monitoring, automation, and optimisation. As a very first step towards data-wise management, we recommend WRRF managers to:

1. Initiate the automation of meta-data collection by enabling data integration and meta-data storage of sensor maintenance actions (installation date, calibration, cleaning, validation, and verification).

2. Assure availability of basic meta-data for online sensor signals in the same location as the sensor signals (e.g., same database). A very basic set of meta-data consists of:
    a. Unit of measurement
    b. Measurement range
    c. Measurement resolution
    d. Measurement principle
    e. Sensor location

3. Prepare for advanced meta-data practices, including the provision of complete historical records of:
    a. The roles and/or purposes of the sensor
    b. History of measured values of sensor offset, sensitivity, trueness, precision, and response time
    c. History of operational state (operational, calibration, validation, maintenance)
    d. History of maintenance protocols for sensor calibration and validation

4. Evaluate the potential of any type of meta-data to prevent expiration of precious data and turning what are currently data graveyards into valuable resources for decision-making.

To stay up to date with the outcomes of the Task Group on Meta-Data Collection and Organisatin (MetaCO) and to learn more on the STR, join the *IWA Connect group*. To learn more about the topic, watch the recording of the IWA webinar *'From data graveyards to data mines'*.

Acknowledgements

The Authors gratefully acknowledge insightful suggestions by Prof Vladan Babovic and Prof Zoran Kapelan. The Authors would also like to thank the IWA Secretariat for support and Vivian Langmaack for taking on the challenge of shaping the paper as it is visually.

References

[1] Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. 2013 *IEEE International Conference on Collaboration Technologies and Systems (CTS)*, 42-47.

[2] Normandeau, K. (2013). Beyond volume, variety and velocity is the issue of big data veracity. Inside Big Data.

[3] Ebbers, M., Abdel-Gayed, A., Budhi, V. B., Dolot, F., Kamat, V., Picone, R., & Trevelin, J. (2013). *Addressing Data Volume, Velocity, and Variety with IBM InfoSphere Streams* V3. 0. IBM Redbooks.

[4] Hauduc, H., Gillot, S., Rieger, L., Ohtsuki, T., Shaw, A., Takács, I., & Winkler, S. (2009). Activated sludge modelling in practice: an international survey. Water Science and Technology, 60(8), 1943-1951.

[5] Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, 21(1), 1-24.

[6] Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P. A., & Comeau, Y. (2010). Data reconciliation for wastewater treatment plant simulation studies—planning for high-quality data and typical sources of errors. Water environment research, 82(5), 426-433.

[7] Campisano A; Cabot Ple J; Muschalla D; Pleau M; Vanrolleghem PA (2013). Potential and limitations of modern equipment for real time control of urban wastewater systems. Urban Water Journal, 10(5), 300-311.

[8] ISO (2003). ISO15839: Water quality – On-line sensors/ Analysing equipment for water – Specifications and performance tests. ISO, Geneva, Switzerland.

[9] Ohmura, K., Thürlimann, C. M., Kipf, M., Carbajal, J. P., Villez, K. (2019). Characterizing long-term wear and tear of ion-selective pH sensors. Water Science and Technology, 80(3), 541-550.

ABOUT THE INTERNATIONAL WATER ASSOCIATION

The International Water Association (IWA) is the leading network and global knowledge hub for water professionals, and anyone committed to the future of water. IWA, which is a non-profit organisation, has a legacy of over 70 years.

IWA connects water professionals in over 130 countries to find solutions to global water challenges as part of a broader sustainability agenda. IWA connects scientists with professionals and communities so that pioneering research provides sustainable solutions.

In addition, the association promotes and supports technological innovation and best practices through international frameworks and standards. Through projects, events, and publications, IWA engages with its members to stimulate innovative ideas and content in support of IWA's vision of a water-wise world.

INTERNATIONAL WATER ASSOCIATION

Export Building, 1st Floor
2 Clove Crescent
London E14 2BE
United Kingdom
Tel: +44 207 654 5500
Fax: +44 207 654 5555
E-mail: water@iwahq.org

Company registered in England No.3597005
Registered Charity in England No.1076690

www.iwa-network.org