https://doi.org/10.71573/py4f7005

A Blind Dive into the Unknown: Water Quality without Metadata

Vincent Pons[1,2,*] https://orcid.org/0000-0001-8574-5674, *Kefeng Zhang*[3], *Luca Vezzaro*[4] https://orcid.org/0000-0001-6344-7131, *Helene Österlund*[1] https://orcid.org/0000-0002-4732-7348, *Tone Merete Muthanna*[1,2] https://orcid.org/0000-0002-4438-2202, *Viviane Furrer*[5,6] https://orcid.org/0000-0002-6993-4958, *Lena Mutzner*[5] https://orcid.org/0000-0002-6954-8360

[1]*Department of Civil, Environmental, and Natural Resources Engineering, Luleå University of Technology, 97187 Luleå, Sweden*
[2] *Department of Civil and Environmental Engineering, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway*
[3] *Water Research Centre, School of Civil and Environmental Engineering, University of New South Wales (UNSW), High St, Kensington, NSW, 2052, Australia*
[4] *Technical University of Denmark, Department of Environmental and Resources Engineering (DTU Sustain), Bygningstorvet, Building 115, 2800 Kongens Lyngby*
[5] *Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland.*
[6] *Institute of Civil, Environmental and Geomatic Engineering, ETH Zürich, 8093 Zurich, Switzerland.*
**Corresponding author email: vincent.pons@ntnu.no*

Abstract

The need to monitor urban water quality is increasing due to the ecotoxicological and health risks urban contaminants pose to water resources. Moreover, the list of contaminants of potential concern is increasing, amplifying the challenges posed by the multiple urban water matrices and sites to be analysed by various analytical methods. Without shared efforts towards a common standardization of datasets and metadata, these monitoring activities will result in datasets fragmented in space, time and context. We propose a first draft for standardizing metadata in urban water quality monitoring, initiating a discussion among the urban drainage community to increase the potential for dataset reuse. Also, this standard will contribute to support the efforts of early career researchers, who are often the main actors in data collection. The proposed data format will allow (re)connecting datasets to generate new knowledge, favour reanalysis and bridge the gap between hydroinformatics and cheminformatics.

Highlights

- FAIR principles in collected urban water quality data
- Why do we need urban water quality metadata
- A draft for standardised water quality metadata structure for review

Introduction

The open science policy led to the emergence of the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) to ensure the quality of published datasets (Wilkinson et al., 2016). Among the FAIR principles, reusability of datasets allows subjects who have not participated in the data collection and publication to benefit from the created knowledge. While reusability is a challenge for most scientific fields, it is especially relevant for urban water quality datasets. Despite the

increasing monitoring activities and capacities over the last decades, water quality datasets remain sparse due to i) the increasing number of compounds to be analysed (contaminants of emerging concern), ii) the large number of sites to be monitored as well as the inter-site variabilities, and iii) the resources needed for field sampling and sample analysis. Hence, single monitoring campaigns are often limited and only combining several datasets will allow the entire field to improve the understanding of the processes driving water quality in urban water systems, eventually leading to better model structures (based on both AI methods and process knowledge).

For instance, the studies by Mutzner et al. (2022) and Zhang et al. (2024) compiled a large amount of micropollutant datasets produced by researchers over the last decade. Those datasets followed different practices in documentation, which made their use challenging (Mutzner, 2024). While the researcher communities generally accept the FAIR principles, these need to be adapted to different scientific fields, and efforts must be carried out to support a wider implementation and adoption in practice (*e.g.*, Jacobsen et al., 2020). In particular, it was acknowledged in the Young Water Professional workshops at ICUD (van der Werf et al., 2024) that the burden of data collection, validation and publication remains vastly on the shoulders of early career researchers, who are already pressed to increase their publication output and find little support to adopt FAIR principles.

To maximise the knowledge gained from monitoring efforts, we must rethink water quality datasets, and provide tools and standards to support the urban drainage field. This study presents the follow-up of the workshop organized during Novatech 2023 on the future of monitoring emerging contaminants (Vezzaro & Mutzner, 2023) and a first draft of coordinated efforts to implement metadata for urban water quality. This draft is intended for discussion and review among the international urban drainage community.

On the need for water quality metadata

Urban water quality data and metadata represent critical components in managing urban water resources, where pollution poses significant risks to public and environmental health. The lack of uniform data formats for water quality data, formats and sampling methods complicates comparing water quality across different sampling sites and urban areas. It hinders the efficient development of national and international regulation and mitigation measures, ultimately impacting environmental and public health protection. Moreover, to improve and guide an optimal design of new monitoring campaigns can benefit from information on previously collected datasets, and on their context (i.e., information about the monitoring objectives, knowledge about the site, methods used, and choices made).

Figure 1 Illustrates an example of a data collection process for water quality in a fictitious catchment over time. Some data are collected at the CSO (Combined Sewer Overflows) of site A in projects 1 and 3. However, between the two projects, time has passed, and the site has changed. It is important to communicate to the user the changes that have occurred, both in the catchment and in terms of the adopted monitoring approach used.

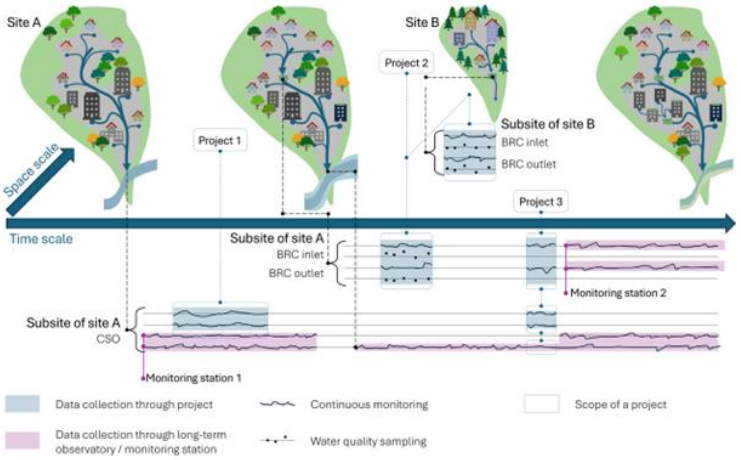

Figure 1: Conceptual view of the evolution of monitoring sites through time

Decades ago, Harremoës (1988) called for accepting the unexplained variability as an inherent and irreducible part of water quality data. In the current era of big data and rapidly developing AI tools, we argue that increased access to shared data across multiple sites, promoted by a fundamental shift in monitoring and documenting datasets, will enable us to cast a new light on several water quality processes. This will create a new understanding of water quality drivers through a data-centric approach to hydroinformatics (Zolghadr-Asli et al., 2024) in which the role of metadata even more crucial for effective water quality monitoring.

Metadata Structure

Metadata provides essential context for the interpretation and analysis of collected water quality data. They include information about the origin, methodology, and characteristics behind each measurement, facilitating comparisons across different geographic sites, temporal scales and contaminants, and allowing researchers to identify trends and make informed decisions regarding water management. We suggest here a first draft describing the collected water quality data types (data values) and required metadata information. The suggested outline (Figure 2) has been developed based on experience from other water sectors (*e.g.*, Therrien et al., 2020), a workshop at Novatech, Lyon 2023, and expert feedback. The expert feedback and discussion will be extended as part of the activities of the JCUD International Working Group on Emerging Contaminants. The following five main metadata categories have been defined:

- WHAT - Variables: descriptive information about what variables have been measured
- WHERE - Locations: spatial location and catchment characterises
- HOW - Methods: information on the sampling, laboratory analysis and reporting methods
- LOG – Occurrences: Logbook describing occurrences and actions performed
- FIND IT – Availability: Lists information about the dataset

Currently, the sub-categories for these five metadata categories in Figure 2 are mainly based on the collection of chemical water quality data. The next step will further develop these subcategories to cover other data types (see Section Testing of metadata).

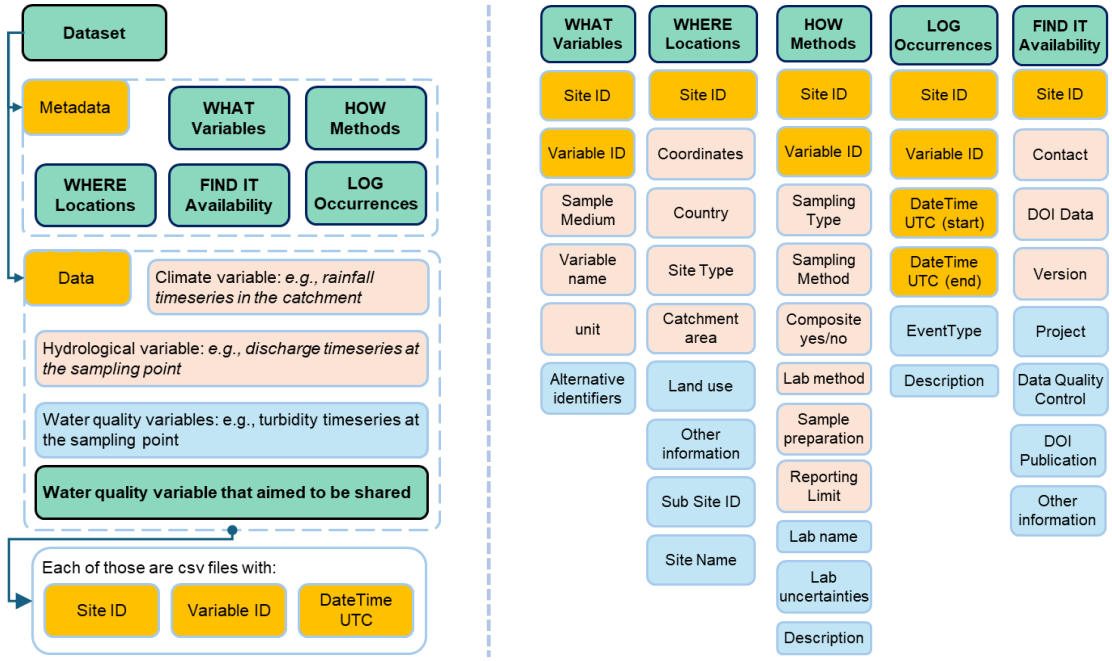

Figure 2: Draft (for discussion and review) for a structure of water quality datasets and metadata. Yellow boxes indicate a unique ID linking the metadata with the collected data values. Light orange boxes indicate necessary metadata, and blue boxes indicate recommended metadata.

Testing of metadata

The proposed metadata format will be tested with potential end-users, including researchers and especially early career researchers. It will be tested on several data types, including micropollutants, results from effect-based methods, online water quality sensors, water quality datasets from blue-green infrastructures, etc. The datasets will be documented in collaboration with several international research groups to adapt the metadata format and increase its adoption.

Future perspectives

We present a structured metadata format draft for urban water quality data. The aim is to allow the inter-site and time comparison of collected data to improve water quality management, water resources protection and regulation. The current draft will be revised in discussion with the urban drainage community at UDM and as part of the activities of the JCUD International Working Group on Emerging Contaminants.

Acknowledgements

The authors would like to thank all the participants of the Novatech 2023 workshop, whose feedback and input influenced the current first draft for standardised metadata. We hope this collaboration will continue, with the urban drainage community providing critical feedback and input in the future.
VP, HÖ and TMM would like to acknowledge the Vinnova competence center DRIZZLE (grant 202203092). The Horizon Europe research and innovation program provided financing to VP and TMM through the project StopUP (grant 101060428), and to LV and LM through the project URBAN M2O (grant 101180710). URBAN M2O is also financed by the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

Harremoës, P. (1988). Stochastic models for estimation of extreme pollution from urban runoff. *Water Research*, *22*(8), 1017–1026. https://doi.org/10.1016/0043-1354(88)90149-2

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, *2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

Mutzner, L. (2024, June 12). *Emerging contaminants in Urban Drainage* [Keynote lecture]. International Conference on Urban Drainage 2024, Delft, the Netherland. https://program-icud2024.iwcconferences.com/event/4777/

Mutzner, L., Furrer, V., Castebrunet, H., Dittmer, U., Fuchs, S., Gernjak, W., Gromaire, M.-C., Matzinger, A., Mikkelsen, P. S., Selbig, W. R., & Vezzaro, L. (2022). A decade of monitoring micropollutants in urban wet-weather flows: What did we learn? *Water Research*, *223*, 118968. https://doi.org/10.1016/j.watres.2022.118968

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 Science and Technology*, *82*(12), 2613–2634. https://doi.org/10.2166/wst.2020.393

van der Werf, J., Lechevallier, P., Smyth, K., Pons, V., & Shi, B. (2024, June 10). *Young Water Professionals' Perspectives on the Future of Urban Drainage Academia* [Conference workshop]. International Conference on Urban Drainage 2024, Delft, the Netherland. https://program-icud2024.iwcconferences.com/event/4770/

Vezzaro, L., & Mutzner, L. (2023, July 3). *Designing the water quality monitoring of the future* [Workshop]. Novatech 2023, Lyon, France. https://website-12399.eventmaker.io/en/programme/642402f8409a420517f4628c

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.18

Zhang, K., Zheng, Z., Mutzner, L., Shi, B., McCarthy, D., Le-Clech, P., Khan, S., Fletcher, T. D., Hancock, M., & Deletic, A. (2024). Review of trace organic chemicals in urban stormwater: Concentrations, distributions, risks, and drivers. *Water Research*, *258*, 121782. https://doi.org/10.1016/j.watres.2024.121782

Zolghadr-Asli, B., Ferdowsi, A., & Savić, D. (2024). A call for a fundamental shift from model-centric to data-centric approaches in hydroinformatics. *Cambridge Prisms: Water*, *2*, e7. https://doi.org/10.1017/wat.2024.5