How reusable are your data? - Towards truly FAIR open data for urban drainage

J. Rieckermann[1*], P. Lechevallier[1,2], J. Agustsson[3], L. Rossi[3], S. Tait[4]

[1]*Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland*
[2]*ETH Zürich, Institute of Environmental Engineering, 8093 Zürich, Switzerland*
[3]*SINEF S.A., Rte des Fluides 1, 1762 Givisiez*
[4] *University of Sheffield, S1 3JD, Sheffield, UK.*
**Corresponding author email: joerg.rieckermann@eawag.ch*

Highlights

- We critically review some examples of urban drainage Open Research Data
- Re-using non-contact spectral pollution monitoring data slightly improved the predictive performance through more reliable data analysis
- Bottlenecks are mostly concerned with the interoperability and the reusability of the data

Available, reliable, and reusable datasets from sewers would have enormous value

Generating new insight from existing data is a major cornerstone of the scientific process. Re-using existing observations with a fresh idea, possibly including complementary datasets, may answer questions that the initial investigators did not even consider. Or demonstrating that an existing idea can be usefully applied to a wider range of circumstance can deliver significant benefits to society. In this fashion, Open Research Data are at the heart of scientific discovery and open science ("Research Parasite Award," 2021). As we all know, experimentation with wastewater and especially in full-scale urban drainage systems is laborious, potentially hazardous and often costly. Consequently, a culture of open (research) data in the field of urban drainage could save substantial costs and reduce project risk especially for research and industrial R&D. However, in the urban drainage area only very few open research datasets exist, which can be difficult to interpret and hence reuse (Abdel-Aal et al., 2018; Nedergaard Pedersen et al., 2021; NYS, 2013; Rivertrust, 2022; Špačková et al., 2021).

Reusing observations in our field, is challenging, among other things, because observations come from various sensors, such as weather radars, ultrasonic depth and velocity probes, or spectrometer probes, for which detailed meta-data, e.g. sensor failures (Figure 1, left) are lacking. Thus, it can be difficult to interpret the data or to assess their "usability" in other studies (Figure 1, right). Openly available, reliable, and re-usable datasets would make it possible to i) study domain-specific processes, ii) build exploratory and predictive models, iii) better calibrate existing models, vi) assess the impact of policies, etc. Several funder of research now encourage researchers to follow the FAIR principles (Finable, Accessible, Interoperable and Re-useable) when collecting project specific data (EC, 2022). Unfortunately, in our field, generally accepted standards are lacking.

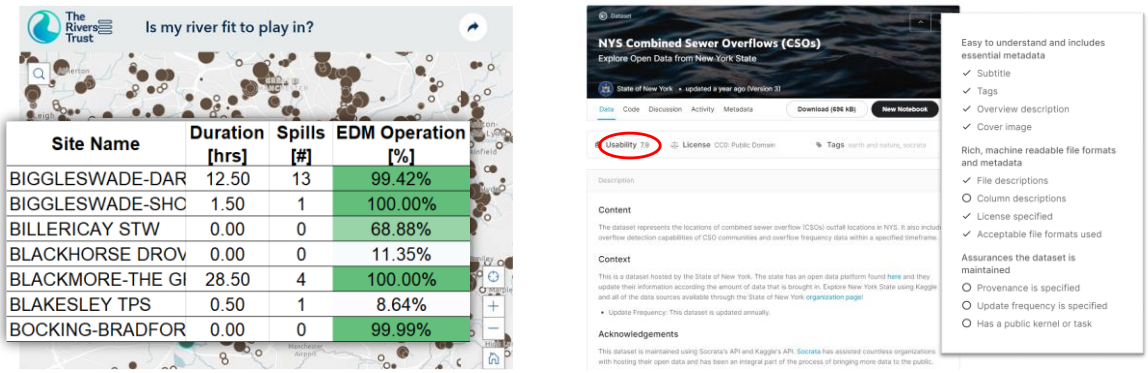

Figure 1. Left: Openly available dataset on CSOs in the UK, which is provided in .csv format, has good metadata and thus is INTEROPERABLE and REUSABLE. Right: Open CSO data from the State of New York, which has a comparably high "usability" score, due to specified licenses for re-use, and meta-data descriptions. Formal assurances that this dataset is maintained are lacking.

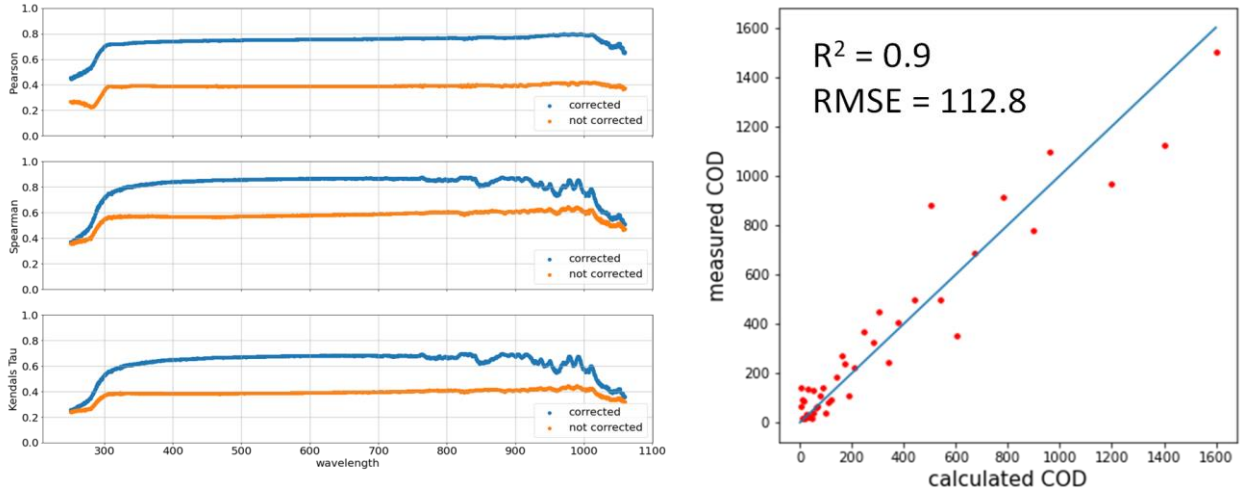

Figure 2. The reuse of spectroscopy data improves the predictive power. Left: Re-analysis (blue) leads to higher correlation coefficients and subsequently to better predictions of COD (right) compared to the original study. While bottlenecks of reusing the data concern all FAIR aspects, most issues in this study concerned **I**NTEROPERABILITY and **R**EUSEABILITY.

In this paper, we present our experience with reusing a unique research dataset on non-contact monitoring of wastewater pollution and then discuss issues regarding implementing of the FAIR principles in the urban drainage community.

Predicting COD from non-contact monitoring of wastewater pollution

In a ground-breaking study, Agustsson et al. (2014) used non-contact monitoring to predict simultaneously the chemical oxygen demand (COD) and turbidity (TUR) concentrations from diffuse reflectance UV-Vis spectra (200–1100 nm). The measured spectra were analyzed using partial-least-squares (PLS) regression. The correlation coefficient between the measured and the reference concentrations of COD was promising (R^2 =0.85). In our re-analysis of the data, we improved the predictive power for COD (R^2 =0.90) slightly, by applying the following steps: First, we identified relevant wavelengths by correlation analysis (Figure 2, left, blue lines). Second, we repeated the calculations, considering the additional COD of the NTU turbidity standard (0.45 (mgO2/L)/NTU) (red lines). Third, we re-fitted their original leave-one-out PLS model (Figure 2, right).

Improving the reusability needs community standards and incentives

Regarding the FAIR principles for open data, we draw the following conclusions: To FIND the data, we had to contact the authors, who provided them by email. While they were ACCESSIBLE, because the data were saved in the commonly accepted ASCII standard, they were not machine-readable without the need for specialized or ad hoc algorithms for translating and mapping. Thus, INTEROPERABILITY and REUSABILITY were only achieved through several in-depth conversations with the data producers. In our view, this was mostly, because accepted standards are lacking in our community. At the SPN10 conference, we would like to discuss how i) to develop a culture in which data interoperability can be expected and ii) we improve incentives to reuse data and so improve the quality and impact of research from the sewer processes and networks community.

Acknowledgements

This work is partly supported by the EU's H2020 research and innovation programme grant no. 101008626.

References

Abdel-Aal, M., Shepherd, W., Shucksmith, J., Tait, S., 2018. CENTAUR project laboratory testing data. https://doi.org/10.5281/zenodo.1406296

Agustsson, J., Akermann, O., Barry, D.A., Rossi, L., 2014. Non-contact assessment of COD and turbidity concentrations in water using diffuse reflectance UV-Vis spectroscopy. Environ. Sci. Process. Impacts 16, 1897. https://doi.org/10.1039/C3EM00707C

EC, 2022. Data Guidelines [WWW Document]. Open Res. Eur. - Data Guidel. URL https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#fairdata (accessed 3.31.22).

Nedergaard Pedersen, A., Wied Pedersen, J., Vigueras-Rodriguez, A., Brink-Kjær, A., Borup, M., Steen Mikkelsen, P., 2021. The Bellinge data set: open data and models for community-wide urban drainage systems research. Earth Syst. Sci. Data 13, 4779–4798. https://doi.org/10.5194/essd-13-4779-2021

NYS, 2013. NYS Combined Sewer Overflows (CSOs) [WWW Document]. Kaggle. URL https://kaggle.com/datasets/new-york-state/nys-combined-sewer-overflows-csos (accessed 3.24.22).

Prodanović, D., Branisavljević, N., 2021. Data archiving and meta-data – saving the data for future use, in Metrology in Urban Drainage and Stormwater Management: Plug and Pray. IWA Publishing, pp. 391–413.

Research Parasite Award, 2021. . Wikipedia.

Rivertrust, 2022. Is my river fit to play in? [WWW Document]. URL https://experience.arcgis.com/experience/e834e261b53740eba2fe6736e37bbc7b (accessed 3.14.22).

Špačková, A., Bareš, V., Fencl, M., Schleiss, M., Jaffrain, J., Berne, A., Rieckermann, J., 2021. One year of attenuation data from a commercial dual-polarized duplex microwave link with concurrent disdrometer, rain gauge, and weather observations. Earth Syst. Sci. Data Discuss. 2021, 1–26. https://doi.org/10.5194/essd-2021-3