Co-UDlabs

Building Collaborative Urban Drainage research Labs communities

D2.1 Report on Data Harmonization

Date of delivery - 31/10/2023

J. Rieckermann, A. Chavarría (EAWAG); J.L. Bertrand-Krajewski (INSA); J. Anta (UDC); A. Schellart (USFD); T. Bruggemann (IKT)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°101008626

DOCUMENT TRACKS DETAILS

Project acronym	Co-UDlabs
Project title	Building Collaborative Urban Drainage research labs communities
Starting date	01.05.2021
Duration	48 months
Call identifier	H2020-INFRAIA-2020-1
Grant Agreement No	101008626

Deliverable Information	
Deliverable number	D2.1
Work package number	2
Deliverable title	D2.1 Report on Data Harmonization
Lead beneficiary	Eawag
Author	**Eawag:** Jörg Rieckermann Alfredo Chavarría **INSA Lyon:** Jean-Luc Bertrand-Krajewski **University of Sheffield:** Alma Schellart **IKT:** Thomas Brüggemann **University of A Coruña (UDC):** Jose Anta
Due date	31/10/2023
Actual submission date	31/10/2023
Type of deliverable	Report
Dissemination level	Public

VERSION MANAGEMENT

Revision history and quality check			
Version	Name	Date	Comment
V 0.1	Jörg Rieckermann (Eawag)	26/04/2023	First draft
V 0.2	Jörg Rieckermann (Eawag), Alfredo Chavarría (Eawag)	18/09/2023	Main results
V 0.3	Thomas Brüggemann (IKT), Yongxiang Qiao (IKT), Simon Tait (USFD)	24/10/2023	Harmonization of CCTV coding
V 0.4	Alma Schellart, Sion Tait (USFD)	25/10/2023	Revision of the documents
V 0.5	Jose Anta (UDC)	27/10/2023	Revision of the document
V1.0	Alfredo Chavarría (Eawag)	31/10/2023	Final version

TABLE OF CONTENTS

List of tables

List of figures

Background: about the Co-UDlabs Project

Co-UDlabs is an EU-funded project aiming to integrate research and innovation activities in the field of Urban Drainage Systems (UDS) to address pressing public health, flood risks and environmental challenges.

Bringing together 17 unique research facilities, Co-UDlabs offers training and free access to a wide range of high-level scientific instruments, smart monitoring technologies and digital water analysis tools for advancing knowledge and innovation in Urban drainage systems.

Co-UDlabs aims to create an urban drainage large-scale facilities network to provide opportunities for monitoring water quality, UDS performance and smart and open data approaches.

The main objective of the project is to provide a transnational multidisciplinary collaborative research infrastructure that will allow stakeholders, academic researchers, and innovators in the urban drainage water sector to come together, share ideas, co-produce project concepts and then benefit from access to top-class research infrastructures to develop, improve and demonstrate those concepts, thereby building a collaborative European Urban Drainage innovation community.

The initiative will facilitate the uptake of innovation in traditional buried pipe systems and newer green-blue infrastructure, with a focus on increasing the understanding of asset deterioration and improving system resilience.

List of acronyms

Acronym / Abbreviation	Meaning / Full text
CA	Consortium Agreement
GA	Grant Agreement
JRA	Joint Research Activity
RI	Research Infrastructure
TA	Transnational Access
UDS	Urban Drainage Systems
IWGDM	International Working Group on Data and Models
OGC	Open Geospatial Consortium
DMP	Data Management Plan
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interpretable, Re-usable
XML	Extensible Markup Language
JSON	JavaScript Object Notation
XLSX	Microsoft Excel
CSV	Comma-Separated Values
GIS	Geographic Information Systems
SI	International System of Units
UCUM	Unified Code for Units of Measure
CEN	European Committee for Standardization
CC0	Creative Commons Zero
CC BY	Creative Commons Attribution
O&M	Observations and Measurements
UDMT	Urban Drainage Metrology Toolbox

Executive summary

This document is a deliverable of the Co-UDlabs project, funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No 101008626. This deliverable is an output from Task 2.1 of Work Package 2 (WP2) "Harmonisation and Capacity-Building".

In addressing these challenges, WP2: D2.1 has delineated three primary objectives: assessing harmonization requirements within the consortium, exploring solutions to identified challenges, and evaluating the broader urban drainage community's harmonization needs. Consequently, three strategic goals have been formed. First, the establishment of core governance and selection of a data sharing infrastructure, such as Zenodo. Second, standardizing data practices to boost data interoperability. Lastly, the promotion of best practices, using standards for data representation and exchange.

Highlighting the progress made during the initial reporting period, a core consortium has been formed to guide harmonization efforts, meeting monthly to assess progress. Zenodo has been chosen as a primary platform for data sharing due to its robust features. The Data Management Plan (DMP) underwent revisions based on the findings from Transnational Access (TA) and Joint Research Activities (JRA), revealing areas that needed change.

Some prominent challenges in harmonization are addressed in the deliverable. Data from diverse sources, differences in units of measurement, variable standards across countries, and unique methodologies adopted by individual RIs led to a complex situation. Some issues surrounding data harmonization in urban drainage system inspection, particularly the standardization and interoperability of CCTV-coding in Europe and the challenges posed by terminological inconsistencies.

The deliverable gives a comprehensive overview on the use of Renku platform for data reproducibility, the role of OGC standards in achieving data harmonization, especially in the field of hydrology, and the introduction of the Urban Drainage Metrology Toolbox (UDMT) to facilitate the adoption of metrological best practices in urban drainage. Another aspect is the initiative's integration with the European Open Science Cloud (EOSC), aiming to create a unified data-sharing approach within the urban drainage community, potentially under affiliations like IWGDM or IWA/IAHR.

On a broader perspective, the goals and objectives of Co-UDlabs are in line with Horizon Europe's overarching vision. Creating the basis for a unified data-sharing approach, possibly within the urban drainage community, clearly fosters sustainability and re-using existing data in the future supports the EU's ambitions in research and innovation. Moreover, our activities acknowledge the importance of cost efficiency, multidisciplinary collaboration, and the principles of open science prevalent within the EU research framework.

1. Ensuring interoperability by defining common standards, protocols and methods

1.1. Effective collaboration in Co-UDlabs needs a framework to harmonize sensors, technologies and data transformation procedures

The Co-UDlabs project aims to build a collaborative European Urban Drainage community by providing access to 17 leading European research facilities. Each research facility has its own unique approach to data collection and management which can make it challenging to share data and collaborate effectively.

In the scientific process, experimental infrastructures make it possible to break new ground and, through the collected experimental data, improve our understanding of the functioning of urban drainage infrastructure. Also, the datasets collected in the joint research activities and transnational access missions also have value in the future, e.g., through re-analysis with better models. Therefore, FAIRifying the data collected in the JRA and TA is an important task.

Unfortunately, the lack of harmonization in urban drainage research can make it difficult to deliver FAIR data, therefore, additional actions are needed to compare data and findings from the Co-UDlabs infrastructures, as well as external data, such as the Bellinge dataset which includes several sensors from this Danish urban catchment (Nedergaard Pedersen et al., 2021). This issue can lead to duplication of research efforts, a lack of trust between researchers, and unreliable findings.

Harmonization serves as the means to ensure that data originating from various sources can be effectively assessed within a unified framework. By establishing this harmonization, researchers can have confidence that they are making comparisons using consistent quantities and units. This, in turn, diminishes the redundancy of efforts, sparing researchers the need to invest valuable time in reconciling or comprehending disparate data formats and procedures. Notably, the requirements for data harmonization differ between laboratory-scale, e.g., our facilities, BENS, RTC_rig, HALL, and real-world site labs, such as UWO and FREJLEV, and certainly, within the realm of water utilities, e.g., FLOWBRU and Event-Duration-Monitoring and mapping.

Considering that Europe's Urban Drainage Systems infrastructure boasts an estimated value of €2.5 trillion and confronts imminent challenges such as aging infrastructure, climate change, and population growth, the impetus for harmonization in urban drainage research extends beyond purely academic interests.

In Table 1, we summarize three well-known instances of data harmonization which demonstrate the profound economic and societal significance of standardization. The initial two, the Mars Climate Orbiter ("Mars Climate Orbiter," 2023) and the Hochrhein Bridge ("Hochrheinbrücke," 2023), are often viewed as cautionary tales due to their adverse outcomes, resulting in mission failure and significant resource wastage. In contrast, Covid Data Sharing serves as a remarkable success story, notwithstanding its challenges, as it facilitated a synchronized global response to the pandemic (Tacconelli et al., 2022). Furthermore, large-scale global challenges, such as climate change, highlight the importance of data harmonization. The Intergovernmental Panel on Climate Change (IPCC) works on combining data sources for climate modelling and predictions, emphasizing the ongoing efforts in data harmonization (Gutiérrez et al., 2021).

Another relevant example closer to our domain is the UK Flood Forecasting Centre. Prior to 2007, river flow data collection was managed by the Environment Agency (EA) while the weather data was collected by the Met Office. Following a flood event in June 2007, it highlighted the need for better communication and data integration, leading the government to initiate the Pitt review, and later to the establishment of the Flood Forecasting Centre. Here, data and models from both organizations were harmonized for improved flood forecasting (*Flood Forecasting Centre*, 2023).

Table 1. Examples of Data Harmonization: Failures and Successes.

Aspect	Mars Climate Orbiter Failure	Hochrhein Bridge Planning Error	Covid Data Sharing Success	IPCC on Climate Change	UK Flood Forecasting Centre
Incident Description	Mission failure due to unit conversion error	Planning error during bridge construction	Global data collaboration during a pandemic	Global effort to harmonize climate data	Unpredicted flooding
Date of Incident	36426	Occurred during construction	Ongoing (since 2019)	Ongoing	June 2007
Primary Cause	Unit conversion error	Height reference planning error	Successful global data sharing	Variety data sources and methods	Lack of integrated data between offices
Impact	Loss of spacecraft	Height misalignment of bridge	Coordinated pandemic response	Inaccurate climate predictions and policy recommendatio ns	Loss of human lives, economic loss, property damages
Key Lessons Learned	Importance of precise units and communication	Importance of meticulous planning	Significance of data harmonization	Significance of data harmonization	Importance of data integration
Mitigation Strategies	Improved unit conversion procedures	Adjusted construction plans	Enhanced data sharing platforms	IPCC creation	Flood Forecasting Centre creation
Resulting Changes in Practices	Emphasis on communication and documentation	Focus on international coordination	Strengthened global health preparedness	Data harmonization for climate modelling	Integration of data and models
Ongoing Relevance to Industry	Highlighted the need for quality control	Highlighted planning and measurement standards	Ongoing importance of data sharing in public health	Standardize data in addressing global challenges	Inter-agency collaboration
Public Awareness and Impact	Public attention due to space mission failure	Local legend in the region	Global awareness of data sharing's importance	Increased understanding in climate change data	Awareness of flood risks and need for better responses
Legal and Regulatory Implications	None significant	None significant	Evolving data sharing regulations and policies	Climate related policies globally	Government inquiries like the Pitt review

Harmonization not only concerns using the same units and data formats, but also the same reference system, terminology, and workflows, as in the Covid Data sharing case. A common framework can ensure data from different sources to be evaluated on a common platform, can reduce the redundancy of efforts, and improve the reliability of findings, can facilitate collaborations, and can reduce costs and consumption of resources. Unfortunately, important things are lacking in the Urban Drainage community to have truly FAIR data: i) Standard terminology, ii) Object Identifiers of processes or infrastructures, iii) Common explicit specification of a shared conceptualization, such as ontologies or data models, iv) Common data sharing platform/API.

1.2. Objectives and scope of the harmonization efforts

Within WP2, we have outlined three primary objectives:

O1: Assessment of Harmonization Requirements within the Consortium: Our first objective is to assess the essential prerequisites for achieving harmonization within our consortium. This involves determining the critical factors necessary to ensure that our data adheres to the FAIR principles, thereby facilitating accessibility and usability. Also, we need to identify a common data sharing platform.

O2: Investigation of Remedies for Identified Challenges: Our second objective revolves around investigating potential solutions for the challenges we uncover. On the one hand, this includes exploring the adoption of standardized protocols, methods, and technologies as remedies for the issues we encounter. On the other hand, this also requires identifying suitable tools to transform data into sharable formats, as well as sharing workflows and data versioning/provenance methods.

O3: Community-Wide Harmonization Assessment: Our third objective extends beyond the consortium and seeks to assess the harmonization needs of the broader urban drainage community, e.g., through coordinating with the IWA/IAHR International Working Group of Data and Models (IWGDM). This broader perspective aims to identify and address harmonization requirements at a community-wide level.

Based on these objectives, we have derived three strategic goals for our harmonization framework:

- **Establish Core Governance and Select a Data Sharing Infrastructure:**

Our initial goal is to establish a dedicated core group to oversee and coordinate harmonization efforts within Co-UDlabs. This group will create guidelines and templates to streamline the harmonization process. Additionally, we will select an appropriate data sharing platform like Zenodo to ensure seamless data sharing within the Co-UDlabs community. This approach may also extend to modelling outputs, enabling their use and reuse in compliance with FAIR principles. This broader application can facilitate tasks such as comparing field data with modelling results, assessing different models, and using model outputs as inputs for other models (e.g., using sewer model outputs as inputs for a WWTP model).

- **Standardize Data Practices and Improve Data Interoperability:**

Our second strategic goal focuses on standardizing data practices to ensure uniform and high-quality data collection, publication, and FAIR archiving. We will review the first generation of Data Management Plans, for the joint research activities and the transnational access program and address issues related to terminology, naming conventions, integration with the European Open Science Cloud (EOSC), and data licensing to enhance data interoperability.

- **Promote Best Practices:**

Our third strategic goal revolves around defining and adopting common standards for data representation and exchange, such as those provided by the Open Geospatial Consortium (OGC). We will identify suitable integration tools and draw insights from successful models, emphasizing reproducibility analysis as a best practice.

These strategic goals will guide our efforts in achieving harmonization and ensuring effective data sharing and collaboration within the Co-UDlabs community. In the following, we structure the report according to the strategic goals, and describe the envisioned steps for the next reporting period.

2. Establishing Core Governance, Refining Data Management and selecting a Data Sharing Infrastructure

This section highlights the main achievements during the initial reporting period, specifically focusing on core governance, the evolution of our Data Management Plan (DMP), and the selection of Zenodo as our primary data repository. For a detailed breakdown, please refer to the main reporting document (Anta, Ciambra, et al., 2022).

2.1. Governance Establishment

As described in the main reporting document (Anta, Ciambra, et al., 2022), we established a core consortium comprising all partners. This core group convenes monthly to discuss progress on harmonization and the other tasks related to WP2, such as capacity building and smart governance. This core consortium maintains a record of participants and significant decisions made during the 10 meetings held so far.

2.2. The Data Management Plan (DMP)

We have crafted a guide for data publication and archiving, drawing inspiration from the Eawag publishing guide and incorporating additional elements for comprehensive coverage. Additionally, we've created a data collection template, which offers a standardized framework for the systematic gathering of data. The main motivation was to get early estimates of data storage volumes, and corresponding quotes from data repositories, or hosting services. As we decided to use Zenodo as a storage, this element is not so important anymore.

Based on the findings derived from Transnational Access (TA) and Joint Research Activities (JRA), it is evident that the data collection template requires a thorough revision. Initially, the Data collection template was fashioned based on pre-existing documents, particularly drawing from the Eawag data collection guide. However, its application within the framework of Co-UDlabs unearthed several issues as stated in Table 2.

For the upcoming TA, we have formulated a set of changes to address these concerns effectively, which are outlined as follows:

Table 2. DMP issues

ID of issue	Issue short description	Proposed solution
1	Emphasis on post-publication data management	Incorporate components in DMP that address pre-publication data management
2	Lack of consistency in DMP topics covered	Inclusion of key DMP components and improvement of current ones (e.g., repetition of questions)

When conceptualized as comprehensive documentation of the data lifecycle, a data management plan is an integral component of the data management quality system throughout the research process. We have identified that our current DMP has a focus on post-publication management of data, with no emphasis on pre-publication management of data. The DMP is a live entity, it should update accordingly all through the duration of the project. Changes to procedures should accumulate in the data management plan documentation.

The second issue is the absence of standardized guidelines for the topics that the DMP should encompass. To address this, we conducted a comparative analysis of various DMP components used in research (Williams et al., 2017) and have proposed improvements to the existing template.

Table 3. Data Management Plan Framework

DMP Component	In current DMP?	DMP Section	Improvement
Project personnel, the duration of their association with the project, their roles, responsibilities, data access, training, other qualifications and identifiers.	Yes	Author Project Information	
Description of all data sources	Yes	Data characterisation	Should be stated that this refers to the origin of data
Data and workflow diagrams	No		Inclusion, as a graphical high-level visualization of data processes
Definition of data elements	Partial	Data characterisation	Should include information about data variables, units of measurement, data categories, metadata, definitions
Planned data model during data collection and processing	No		This could include information about sensors or lab methods for data acquisition
Planned data model for data sharing	Yes	Data characterisation Data publication Data preservation	
Observation and measurement	Partial	Data characterisation Dataset details	Should include sensors used, the sampling methods, and the frequency of data collection.
Data recording	No		Should include information about the registry system used. Naming conventions, file structures, version control.
Data processing	No		Should include how raw data is processed and transformed into a usable format.
Data integration	No		Should include methods for integrating data from multiple sources.
Data quality control definitions and acceptance criteria	No		Should include the criteria and procedures for assessing data quality

Calibration plans	No		Should include documents or procedures taken to assure software and sensors are working properly
Configuration specifications for the project	No		Specific configurations or settings for software and devices used
Validation or testing plans	No		Should include plans for testing data processing pipelines, validation scripts, or software components
Scheduled maintenance	No		Schedules for updates, patches, and equipment checks
Plan for handling unscheduled maintenance	No		Contingency plan for dealing with unexpected maintenance issues or equipment failures (only if needed)
Security plan	No		Access control, encryption and data handling protocols (only if needed)
Data back-up and schedule	No		Should include data backup plan (backups storage and frequency)
Privacy and confidentiality plan	Yes	Restrictions	Type of License or compliance to standards, data anonymization (only if needed)
Project management plan (deliverables, timeline, tracking and reporting plan, and resource estimates)	Partial	Ensuring interoperability	Identify mismatches between project management and resources available.
Data retention, archival, and disposal plan	Yes	Data preservation, Data publication	
Data sharing plan	Yes	Data preservation, Data publication	

2.3. Selection of Zenodo as Data Sharing Infrastructure

Furthermore, we have opted for Zenodo as our designated data sharing infrastructure. This decision aligns with the comprehensive analysis detailed in the Periodic Report, as depicted in the accompanying figure. To facilitate the use of the platform, we also created a guide of how to upload data to Zenodo. It is worth noting that the European Open Science Cloud (EOSC) is now operational, offering an opportunity to establish a Co-UDlabs community or potentially elevate our presence within the broader Urban Drainage community, e.g., make an IWGDM community. We plan to explore these opportunities in detail during the remaining duration of the project.

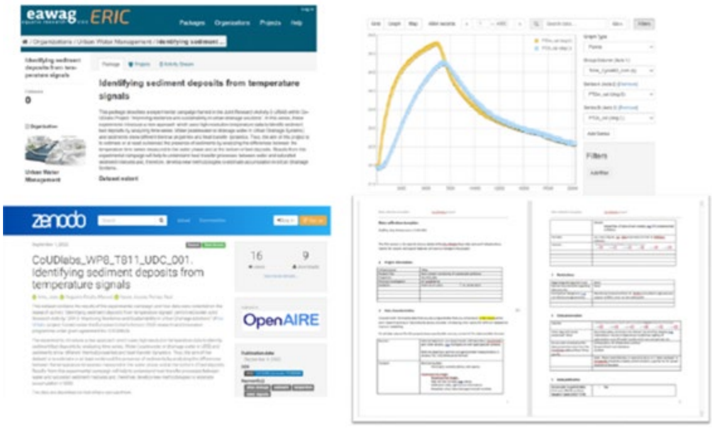

Figure 1: Comparison of Zenodo and ERIC platforms showcasing Co-UDlabs project uploads, and Data Collection Templates analysis (Rieckermann, 2023)

3. Standardize Data Practices and Enhance Data Interoperability

3.1. Review of existing literature related to urban drainage data sharing and FAIR use

While the urban drainage community is small, the involved researchers have always been sharing their data. Notable instances include collaborative efforts in the past to validate hydrodynamic rainfall-runoff models (Fuchs & Harms, 1986; Radojkovic & Maksimovic, 1986). Nonetheless, these endeavours have typically been ad hoc, based on personal contact, and lacking a structured approach. Knowledge and insights have often been consolidated in PDF documents, without a deliberate focus on adhering to FAIR principles for data use and accessibility.

Table 4. Collaborative Data Analysis Efforts within a Common Framework

Task	Description	References
Re-analysis of SCAN Data	Conducting a fresh analysis of SCAN data	(Caradot et al., 2013) and (Lepot et al., 2016)
Water Quality Modeling	Collaborative modeling of water quality involving multiple research groups within the consortium.	(Dotto et al., 2012)

Unfortunately, traditional practice has its limitations. When a replication or reproduction of a study fails, or other questions about how the data was handled arise, traceability and documentation are the only reliable support. For this reason, professional data managers are increasingly concerned about traceability in scientific fields (Williams et al., 2017).

Interestingly, some datasets which are relevant for urban drainage are emerging in the public domain (Table 5). Table 5 describes 12 openly available datasets, which encompass a diverse range of urban drainage and environmental monitoring endeavours. As the licenses for re-use differ substantially, it will be very challenging to create supersets of such data. To make them FAIR and re-usable, data should be published under a CC0 license.

Table 5. Open Datasets UDS

Title	Description	Publications	Dataset	License of the data

Bellinge dataset	Comprehensive dataset for the urban drainage system in Bellinge, Odense, Denmark. Including sensor data, models, background information etc. for a case area during 10 years of observation.	https://essd.copernicus.org/articles/13/4779/2021/	https://doi.org/10.11583/DTU.c.5029124.v1	CC BY 4.0 (rainfall: CC BY-NC 4.0)
COMMON dataset	Dataset of rainfall observations of a dual-polarized microwave link, including DSD measurements	https://essd.copernicus.org/articles/13/4219/2021/	https://zenodo.org/record/4923125	CC BY 4.0
Sewer-ML	A Multi-Label Sewer Defect Classification Dataset and Benchmark	https://arxiv.org/abs/2103.10895		(CC BY-NC-SA 4.0)
UWO opendata	Comprehensive rainfall-runoff dataset in an urban drainage network, including wastewater temperatures and the performance of the wireless telemetry systems, including accompanying data	https://engrxiv.org/preprint/view/3208	Eawag repository	CC0
Micropollutants	A decade of monitoring micropollutants in urban wet-weather flows	https://doi.org/10.1016/j.watres.2022.118968	https://zenodo.org/record/6808401	GNU v3
FLOWBRU	Urban Drainage dataset from The Brussels Capital Region (Belgium)		https://www.flowbru.be/fr	NA, "data can be freely downloaded"
Manhole solute traces	Experimental solute traces (upstream and downstream ttemporal concentration profiles) recorded in model surcharged manholes at USFD	https://ascelibrary.org/doi/10.1061/%28ASCE%29HY.1943-7900.0001951	https://figshare.shef.ac.uk/articles/dataset/University_of_Sheffield_Experimental_Manhole_Traces_and_CRTDs/13373039	
EDM dataset	Event Duration Monitoring (EDM) for CSOs in England and Wales		https://ckan.publishing.service.gov.uk/dataset/event-duration-monitoring-storm-overflows-annualreturns. T	NA
SIPIBEL dataset	Data to investigate hospital effluents characterization, treatability, and impacts within an urban wastewater treatment plant.		https://zenodo.org/record/5176161	

First Flush Dataset	Call for turbidity time series in the "Urban Drainage List" community to analyse first flush phenomena		under construction	
FloodCitiSense project	This FloodCitiSense web app allows you to view rainfall data from official sensor networks, e.g., from the UK Environment Agency, as well as low-cost sensors deployed by citizens. It also allows you to view flood and rainfall reports made by citizens using the FloodCitiSense mobile app.			

In total we identified 12 openly available datasets, which encompass a diverse range of urban drainage and environmental monitoring endeavours. Notable examples include a comprehensive dataset from Bellinge, Denmark, offering a decade's worth of sensor data, models, and background information for in-depth analysis. Additionally, datasets featuring rainfall observations, such as dual-polarized microwave link measurements and drop size distribution data, contribute to our understanding of precipitation patterns.

Further datasets include a multi-label sewer defect classification benchmark, an urban rainfall-runoff dataset encompassing wastewater temperatures and telemetry system performance, and a decade-long record of micropollutants in urban wet-weather flows. Additionally, data from the SIPIBEL Bellecombe pilot site explore hospital effluent characterization within an urban wastewater treatment plant. These datasets collectively offer valuable insights into urban drainage, water quality, and environmental monitoring research.

Three important lessons that we learned from this review are, first, that there is no single repository or community for data sharing. Second, the absence of a CC0 or similarly standardized open license can make data reuse more complex and challenging due to the need to navigate varied and potentially conflicting licensing terms and requirements (see Section 3.6 for a more thorough analysis). Third, data reuse is substantially difficult when file formats are not standardized and standardization of file formats or the use of widely accepted data exchange is clearly needed.

3.2. Data exchange formats

Urban Drainage Systems data covers a range of scientific and technical fields, with data often grouped into datasets. Preserving this grouping is important for maintaining temporal and spatial relationships. So, one early step in planning a data archive is choosing the file formats. This choice impacts robustness, and future usability, including sharing, retrieval, and processing. Our aim is simple: to understand the common file formats used within the community. By identifying adopted formats, we can evaluate whether these formats align with the Co-UDlabs data, promote best practices for data representation, consider data transformation and integration and contribute to long-term preservation of data.

Our approach involves a review of academic literature, repositories, and data exchange infrastructure websites. These steps allowed us to grasp the main formats and data sharing practices within the UDS and wastewater management fields (Table 6).

Table 6. Common Exchange Formats and Corresponding Urban Drainage Datasets

Category	Data Exchange Format	Examples of Observations	Examples datasets
Spreadsheet	CSV XLSX	- Rainfall intensity - Flow rates - water levels - water quality parameters	– Bellinge dataset. – Sustainable Urban Drainage System SUDS Register FCC Dublin.
Markup and Document Format	XML (Extensible Markup Language)	- Weather observations - Water ML 2.0	– NOAA observed weather data. – CUAHSI WaterML.
Data Serialization	JSON (JavaScript Object Notation)	- Real-time sensor data (e.g., IoT sensor data in JSON)	– SEPA Water Level Data (Scotland). – Watershed Water Quality - Wastewater Qualifiers (City of New York).
Scientific Data Format	NetCDF (Network Common Data Form)	- Spatial and temporal rainfall data - Hydrological model outputs (e.g., runoff simulations) - Climate data	– Deltares Global Water Availability
Geographic Information Systems (GIS)	GML Shapefile KML GeoTIFF	- Spatial information of drainage networks - Catchment boundaries - Infrastructure locations (e.g., monitoring sensors) - Land use data - Rainfall data. - Digital Elevation Models (DEM) used in drainage modelling.	– Sewage treatment plants UWWTD - Flanders. – JRC Global Surface Water.
Database File Formats	DBF (dBASE File) Sqlite	Monitoring data.	– UWO dataset.

A noteworthy trend emerged, emphasizing the importance of formats that facilitate effortless interpretation without the need for complex transformations (Prodanović, Dušan & Branisavljević, Nemanja, 2021). Within the academic domain, there is a preference for spreadsheet data formats like CSV (Comma-Separated Values) and XLSX (Microsoft Excel), extensively employed for data exchange. These formats are widely supported for its simplicity, and compatibility. Government entities lean towards using JSON (JavaScript Object Notation) and XML (Extensible Markup Language) formats. XML adheres to interoperability standards supported by OGC, ensuring seamless data sharing across diverse systems.

However, the selection of file formats is not a one-size-fits-all. It is directly linked to the nature and intended use of the data. For datasets with a spatial dimension, Geographic Information Systems (GIS) formats like GeoJSON, GML, and Shapefile, are often preferred. This selection of formats ensures alignment with the specific requirements of the data.

The decision-making process for data format selection within the UDS domain should consider compatibility, interoperability, and data context, as important factors guiding the choices made by both researchers and government entities. This strategic approach streamlines data exchange, promoting effective collaboration and data-driven decision-making within the community.

3.3. Collating and Categorizing Data Management Insights from Co-UDlabs Research Infrastructures

We initiated a series of activities to collect, then analyse, and visualize data management information. The first activity was the distribution of the Data Management Plan (DMP) to all researchers and Research Infrastructure (RI) managers, aiming to collect descriptions of the data that would be gathered. Subsequently, the data was compiled into a structured table.

The analysis process began with extracting data from the submitted DMPs and investigating publicly available information related to the Research Infrastructure. This preliminary step involved the extraction of details, including measurement specifics, units of measurement, and sensor details across all the RIs. We deliberately excluded CCTV coding data, because the harmonization of this data type would go beyond the capacity of the Co-UDlabs project (see Section 3.4).

The data collected from experiments can broadly be categorized into two primary groups: data from observations obtained through sensors and data output derived from laboratory procedures. A considerable portion of the data from Co-UDlabs RIs comes from a range of different types of sensors. In addition to sensor-based observations, other data outputs involve samples collected during laboratory-based experiments and subjected to further analyses. These laboratory-based procedures yield valuable data outputs that provide insights into physical, chemical, and molecular aspects of urban drainage systems and wastewater management. Data resulting from laboratory analyses may include microscopic imagery, molecular analysis results, sediment characteristics, among others.

Recognizing the details and specific models of each sensor, we aggregate them to make the data more manageable. To provide a more structured framework, we introduced a hierarchical classification. This categorization process involved grouping specific sensors into broader, more general functional categories, simplifying the grouping of measurements per sensor.

To maintain detailed sensor information while ensuring data clarity, a description column was added. This column provides supplementary details, including specific sensor models and other relevant specifications.

Definition of Measurement Categories:

We defined distinct measurement categories to categorize the types of measurements used in each RI. These categories were designated to provide a clear overview of the nature and scope of measurements:

- Weather monitoring
- Flowrate and Velocity Measurements
- Microscopic Analysis
- Molecular Methods
- Sediment Analysis
- Structural and Defect Measurements
- Water Level and Depth Measurements
- Water Quality Parameters

The categories provide clarity for the wide array of data generated by the Co-UDlabs Research Infrastructures. The distribution of variables per category and RI is described in Table 7.

Table 7. Distribution of Measured Variables across different Categories and RI

Categories	A/B FLUME	ANNULAR	BENS	BURIED	HALL	IKT LTF	IKT TEST	UWO	Total
Weather Monitoring								6	**6**
Flow and Velocity Measurements	4	4	4	3	1	2	1	2	**22**
Microscopic analysis		4			1				**5**
Molecular methods		4							**4**
Sediment Analysis		6	5	2			1		**13**
Structural and Defect Measurements						3			**3**
Water Level and Depth Measurements	2	1	2	5		1	2	2	**15**
Water Quality Parameters	2	4	9	3	13		3	2	**36**
Total	**8**	**23**	**20**	**13**	**15**	**6**	**7**	**12**	**104**

Figure 2 visualizes the categorization of sensors within Co-UDlabs Research Infrastructures (RIs). The Annular RI stands out as having the most extensive range of categories and variables. Across all RIs measurements are done in the category of 'Flowrate and Velocity'. Additionally, the 'Water Quality Parameters' category shows the greatest variety in sensor types. For a more in-depth analysis and to view specific quantities of each variable access the interactive version of this chart online.

Figure 2: Sunburst chart representing the categorization of measurements within Co-UDlabs RIs. The core shows the RIs, followed by categories and specific variables in the outer layer. An interactive version can be accessed online [link]. (Source: Chavarría, 2023)

3.4. Challenges with harmonizing CCTV-codes in Europe and perspectives for future activities

The optical inspection of drainage systems outside buildings is regulated in DIN EN 13508-2 (DWA, 2014). This European standard establishes a coding system for describing observations of defects (e.g., cracks) and pipe characteristics (e.g., joints, lateral connections) made inside drainage pipes and sewers, manholes, and inspection openings during visual inspection. Visual inspections are normally conducted using CCTV in which the cameras move through pipes and manholes on powered manually controlled platforms. Coding is normally done manually. Nevertheless, there are still national standards with minor deviations from DIN EN 13508-2. There is also variation regarding special additional national standards concretizing DIN EN 13508-2 (DWA, 2014), attributable to historically grown structures. In Germany and France DIN EN 13508-2 is used. In Germany, the EN 13508-2 is supplemented by the DWA M 149-2 standard (DWA, 2013), e.g., regarding the characterization of cracks. In the UK Manual of Sewerage Condition Classification MSCC (WRc, 2013) is used for coding defects. Netherlands and Belgium use the NEN3399 (*NEN 3399*, 2015) to classify and code in-pipe defects. The differences in the national coding are described in detail in Deliverable D07.2 "Assessment of Current Pipe Condition Assessment Approaches and Proposals for Improvement".

Besides European coding according to EN 13508-2 (DWA, 2014), a different approach has been established in the countries over the decades and older standards are still used in parallel (e.g., in Germany: coding according to ATV M 143-2 (ATV, 1999) or according to the ISYBAU standard).

A further standardization for operators goes beyond the current EN 13508-2, therefore does not appear to make sense, because these different kinds of approaches and derivations have been already implemented by many actors (sewer network operators, inspection companies, engineering offices) during inspection, rehabilitation planning and damage repair. Introducing a harmonized standard, that is widely adopted by European Asset Management practice therefore clearly goes beyond the scope of Co-UDlabs.

First, a harmonized coding system is challenging, because the coding must always be seen in the context of the respective question at hand and different purposes encode different aspects. Second, it should be noted in any case in this context that according to Brandl (Brandl & Beschreibung, 2017), despite formal and seemingly objective classification criteria, an assessment is nevertheless still subjective. There is ample evidence that the same damage, assessed by different inspectors, can also yield different classification results (Biere, 2013; Dirksen et al., 2011; DWA, 2005). Therefore, such coding data should be treated with extra caution, e.g., with regard to a systematic scientific evaluation.

In the future, the coding system according to EN 13508-2 could be a starting point for a harmonization activity. It could also be used in principle for scientific questions in the laboratories of the Co-UDlabs partners. Furthermore, it would be useful to start with harmonizing the quality of CCTV-images, e.g., for the further development of artificial intelligence for the detection of defects and their assessment within the framework of research projects. Minimal requirements for CCTV-files are suggested in the German standard (DWA, 2010), e.g., regarding the resolution of pictures, the camera frame and the inspection process. But these requirements may also vary from country to country. In the future, a harmonization of the format of the images and their designation would also be useful. The requirement listed in the German standard DWA M 159-5 (DWA, 2005) could be a basis for discussion on harmonization. Even though the harmonization of the exact method for the collection of CCTV images would start to bring more consistency in the analysis of CCTV inspection of sewers and drainage pipes, it must be realised that the way in which water utilities engage inspection companies is different across Europe. Some water utilities have their own staff and so inspection is effectively a fixed cost and so image quality is normally higher. Other utilities purchase inspection services at a unit rate and this contracting format encourages inspection companies to carry out inspections quickly with less incentive on image quality.

XML specifications are available (e.g., WRc (2013)) to allow the consistent structuring of sewer inspection data into a format that can be automatically handled by asset databases, but these are in general proprietary and require asset databases of the water utility to be flexible in terms of data input. So, no general specification is available that is widely used throughout Europe.

Recent trends in sewer asset inspection data analysis will probably result in less standardization. With the advent of machine learning approaches to CCTV image analysis there have been several water utilities and commercially driven projects to collect, and label CCTV data and then to others to use data-driven to produce proprietary algorithms to automatically identify sewer defects and artifacts. The focus on providing open datasets is to encourage innovative SMEs to enter the sewer inspection market. Commercially they need to protect their know-how so that the likelihood of developing open and standard ways of CCTV image analysis is likely to be low.

3.5. Identification of issues with the used terminology

Following the data harmonization process, we encountered a critical challenge: inconsistent terminology. Diverse facilities use distinct naming standards for variables and sensors, leading to complexities in data interpretation, discovery, and retrieval. This section identifies the inconsistencies related to variables naming, sensor names, unit standards, data collection frequency, metadata standards, and storage needs. These considerations are important for achieving data integration and interoperability among the various datasets generated from the Co-UDlabs TA and JRA.

3.5.1. Naming inconsistencies

The difference in naming standards poses challenges not only in data interpretation but also in data discovery and retrieval. Different facilities refer to the same variables using distinct names, abbreviations, and case-sensitivity differences, making it difficult to determine whether the variables are indeed referring to the same parameters. For example, variables like "flow," "flow rate," "flow rates" or "flow discharge" might all represent similar measurements but are labelled differently across datasets. This variation in terminology complicates data interpretation, data discovery and retrieval, and also prevents the use of automated data analysis. To address this issue, standardizing variable and sensor names is important for integration and interoperability among diverse datasets. The resultant Table 8 provides an overview of the observed differences within each variable group.

Table 8. Standardization of Variable Names and their varieties in Data Collection

Name	Variables names variety
Flow Rate	Flow rates, Flow Rate, Flow Discharge, Flow Rate at Upper Storm Sewer, Overflow Flow Rates, Inlet Flow Rates to TES, Inlet Flow Rates to Perforated Pipes, Flow rate, Flow discharge, Flow
Flow Velocity	Velocity, Local 2D Velocity, Flow velocity, Flow Velocity, Velocity, 2D Velocity, 3D Velocity
Temperature	PT100 probe (water temperature), Temperature sensors (Temperature [ºC]), Temperature probes, Processed temperature time series
Organic Matter	Organic matter, Percent organic matter
Grain Size	Mean grain size, Median grain size [d50]

Conductivity Measurements	Conductivity, Dielectric conductivity
Density Measurements	Wet density, Bulk density, Dry bulk density, Solid density, Density, Density [ρ]

To identify standard variable names, we did a research process focused on Urban Drainage Systems (UDS) glossaries, vocabulary resources, ontologies, and relevant naming conventions or standards. While a dedicated standard specific to the UDS community is not readily available, we compared the following sources to define the common variable names:

- Standard Methods for the Examination of Water and Wastewater
- Urban Drainage: A Multilingual Glossary
- International Glossary of Hydrology
- ISO 772:2022 Hydrometry — Vocabulary and symbols
- ISO 6107:2021 Water quality — Vocabulary
- IEC 81346-2 Industrial systems, installations and equipment and industrial products – Structuring principles and reference designations – Part 2: Classification of objects and codes for classes
- SUSDRAIN Glossary

From this review, we curated a set of variable names in Table 9, which served as a reference for the next stages of our analysis.

Table 9. Comparative Analysis of Variable Nomenclature across multiple sources

Variable standard	S.M.E.W.W.	UD: A Multilingual Glossary	International Glossary of Hydrology	ISO 772:2022 Hydrometry	SUSDRAIN Glossary	IEC 81346-2
Ammonium Nitrogen	Ammonium Nitrogen	Ammonium				
Chemical Oxygen Demand	Chemical Oxygen Demand	Chemical Oxygen Demand	Chemical Oxygen Demand			
Conductivity (*)		Conductivity	Hydraulic, Electrical Conductivity	Hydraulic, Electrical Conductivity		
Density (*)			Density, Relative density	Bulk density		Density
Depth			Depth			
Dissolved Oxygen	Dissolved Oxygen	Dissolved Oxygen	Dissolved Oxygen Content			
Flow Depth		Flow Depth				
Flow Pressure		Pressure, Pressurized flow				

Flow Rate		Flow Rate			Flow	
Flow Velocity		Flow Velocity	Flow Velocity			
Humidity (*)			Atmospheric, Relative, Absolute, Specific Humidity			Humidity
Grain Size (*)		Grain size distribution	Grain size distribution, Grain size	Grain size distribution, Grain size		
Moisture Content (*)		Antecedent moisture	Soil Moisture, Soil Moisture Content			
Organic Matter		Organic Matter	Particulate Organic Matter			
Orthophosphate Phosphorus	Orthophosphate Phosphorus	Orthophosphat e				
ph		pH	pH			
Porosity		Porosity	Porosity	Porosity	Porosity	
Pressure		Pressure				Pressure
Rainfall (*)		Rainfall Runoff-model, time series, Intensity	Rainfall Intensity, Depth, Distribution, Runoff-model	Rainfall, Rainfall intensity		
Temperature		Temperature				Temperat ure
Thermal Conductivity			Thermal Conductivity			
Total Nitrogen	Total Nitrogen	Total Nitrogen				
Total Phosphorus	Total Phosphorus	Total Phosphorus				
Total Solids	Total Solids	Total Solids				
Total Suspended Solids	Total Suspended Solids	Total Suspended Solids	Total Suspended Solids		Total Suspended Solids	
Turbidity	Turbidity	Turbidity	Turbidity		Turbidity	
Water Level		Water level	Water level	Water level		Level

* **Note**: While these terms (e.g., "Rainfall") are often used as general references, the specific context requires a more precise description. Specific terms when narrowed down (e.g., "Rainfall Intensity") can provide a detailed insight into the variable being referred to.

3.5.2. Unit inconsistencies

Diverse units in the datasets can complicate interoperability and data exchange. While standards, such as the International System of Units (SI), Unified Code for Units of Measure (UCUM), and the standards set by the European Committee for Standardization (CEN) exist, not all datasets adhere strictly to these standards. Using a universally recognized and standardized unit system, such as the SI, can ease data sharing and integration processes. Table 10 encompasses the units for all measurements in the RIs:

Table 10. Comparative Analysis of Variable Nomenclature across multiple sources

Variable	Unit	UCUM	SI		Variable	Unit	UCUM	SI
Ammonium Nitrogen	mg/L	✓	✓		Porosity	pu		
COD	mg/L	✓	✓		Pressure	bar	✓	✓
Conductivity	µS/cm, mS/m	✓	✓		q-PCR	Ct		
CRISPR-Cas	DNA, RNA				Rainfall	mm/h	✓	✓
Density	kg/m3	✓	✓		Simulation Damage	[undetermined]		
Depth	mm	✓	✓		Solar Radiation	W/m2	✓	D
Digital PCR	copies/µL				Solute Concentr.	mg/L	✓	✓
Dissolved Oxygen	mg/L	✓	✓		Strain	micrometer/m	✓	✓
Flow Depth	mm	✓	✓		Substance Retention	[undetermined]		
Flow Pressure	mA		D		Synthetic Wastewater Model	[undetermined]		
Flow Rate	L/s, m3/s	✓	✓		Temperature	Celsius	✓	D
Flow Velocity	m/s	✓	✓		Thermal Conductiv.	W/m/Celsius	✓	D
Humidity	m3/m3	✓	✓		Total Nitrogen	mg/L	✓	✓
Imagery	[undetermined]				Total Phosphorus	mg/L	✓	✓
Leak Tightness	[undetermined]				Total Solids	g/Kg, mg/L	✓	✓
Mean Grain Size	mm	✓	✓		TSS	mg/L	✓	✓
Metagenomics	DNA, RNA				Turbidity	NTU	✓	
Microbiological Analysis	[undetermined]				Volumetric Heat Capacity	J/m3/C	✓	D
Moisture Content	%				Vol. Moisture Content	m3/m3	✓	✓
Organic Matter	%				Water Level	mm, m	✓	✓
Orthophosphate Phosphorus	mg/L	✓	✓		Water Infiltration	m3/m3	✓	✓
ph	ph	✓			Water Overflow	L	✓	✓

Physico-chemical Analysis	[undetermined]				Wind	m/s	✓	✓

Most units in the dataset adhere to established standards such as SI or UCUM. However, not all units are explicitly defined, potentially complicating data interpretation. Certain measurements, especially those related to DNA or RNA sequences, Ct values, and those labelled as "undetermined," need a clearer unit description. Table 11 show some variables that exhibit varying units, for instance, conductivity represented as µS/cm or mS/m, and Total Solids denoted as g/Kg or mg/L. For these discrepancies, we need to ascertain whether they are accepted within the context or if there's potential to standardize the units for enhanced harmonization.

Table 11. Variability in Units for Common Variables in the RI

Variable	Water level	Total Solids	Flow Rate	Conductivity
Unit	m, mm	g/Kg, mg/L	L/s, m3/s	Us/m, mS/m

Another issue is related to the frequencies with which the source data is collected, and its variations within the Co-UDlabs facilities. For example, while one facility collects data at a high temporal resolution of every 5 minutes, another may opt for hourly intervals. This discrepancy introduces challenges related to temporal misalignment, potentially leading to data gaps or overlaps when merging or analyzing datasets.

Metadata plays an important role in providing context for collected data. Inconsistent or incomplete metadata can lead to misinterpretations when analysing datasets. Metadata should describe the data, including its content, structure, methodology, and any associated resources.

Different Co-UDlabs facilities may have diverse storage needs for their data, leading to different storage systems. Addressing these storage needs and the data accessibility can help establish guidelines on a common data archiving platform. Co-UDlabs data need to be archived for long-term preservation. Archiving strategies, including storage technologies and metadata for retrieval, become relevant in this context.

3.6. Identification of a suitable license for effective sharing of Co-UDlabs and UDS data

Table 5 (in Section 3.1 above) showed that the available open datasets have very different licenses which regulate how the data can be re-used. This is important, because the status of research data and its eligibility for copyright protection can have significant implications for how it can be shared and re-used.

The inherent nature of data often does not meet the criteria for copyright protection due to its absence of intellectual creation (Nunnenmacher, 2023). As a result, the concept of applying a license to data can seem incongruous or, at best, clumsy (von Waldow, 2017). However, a grey area emerges when considering the presentation or potential intellectual compilation of data. In these instances, where data is skilfully organized, curated, or compiled, there may be grounds for applying copyright protections to the specific presentation or compilation, rather than the underlying data itself.

In the practical realm of data utilization, especially when dealing with comprehensive datasets compiled from numerous sources, the presence of a CC0 (Creative Commons Zero) license proves exceedingly advantageous (Gee, 2023). CC0, effectively waiving copyright, allows for unrestricted use and reusability of data. On the contrary, employing a CC BY (Creative Commons Attribution) license, or other restrictive licenses, can create practical hindrances.

To maximize data re-use in Co-UDlabs, adopting a CC0 license is conducive to fostering accessibility and encouraging unrestricted analysis, particularly when working with complex, multi-sourced datasets. For further details see (Gee, 2023) and Appendix 6.1, which also gives details on the DMLawTool, a practical resource for identifying examples of research data protected by copyright.

4. Promote Best Practices

4.1. Example of knowledge infrastructure/common workflows: Reproducibility analysis

4.1.1. Selecting a platform to support reproducibility

As part of the harmonization efforts and compliance with the FAIR principles, we carry out the task of identifying platforms to support reproducibility. Platforms like Google Colab and Binder have made advances in this direction. Google Colab simplifies the computing environment, notably by enabling zero-configuration execution of Jupyter notebooks. Binder, on the other hand, offers transparency in the execution steps of Jupyter notebooks within a containerized setup, turning any repository URL into a live compute environment in a short time. While these platforms increase project visibility and reproducibility, they lack any means of tracking dataset usage within the projects (Roskar et al., 2023).

Renku stands out as an open-source platform for collaborative data science initiatives (for more details, visit the Renku website). It provides an all-encompassing environment where data scientists can efficiently manage data, track iterations, and reproduce experiments. Renku presents itself as a web platform (RenkuLab) and as a command-line interface (Renku Client). It's constructed atop a foundation of open-source elements, tailored to researchers, data scientists, educators, and students, assisting them in data, code, workflow management, tracking data provenance, and setting up computational environments. While Renku itself doesn't store data directly, it provides features and integration with storage systems, like Zenodo, to manage and work with large datasets effectively. Some key features of Renku include:

- Version Control: Renku integrates with Git for version control, allowing users to track changes, collaborate, and maintain a history of project modifications.
- Data Management: Renku provides tools for data management, enabling users to organize, share, and track data throughout the project lifecycle. It supports data versioning, provenance tracking, and metadata management.
- Workflow Management: Renku helps users design and manage computational workflows. It supports defining dependencies, executing workflows, and capturing intermediate results.
- Collaboration: Renku facilitates collaboration among team members. It allows multiple users to work on the same project simultaneously, enabling seamless integration of their contributions.
- Reproducibility: Renku promotes reproducibility by capturing the environment, dependencies, and execution steps. This enables the reproduction of results and facilitates sharing and verification of research outcomes.
- Integration: Renku integrates with various tools and services commonly used in data science, such as Jupyter Notebooks, RStudio, and containerization technologies like Docker.

4.1.2. Building a reproducible workflow in Renku

We decided to test Renku using a dataset from the Co-UDlabs TA: Co-UDlabs_WP8_T812_UDC_001 (Anta, Regueiro-Picallo, et al., 2022). This dataset, hosted on Zenodo, presents the experimental campaign's results for the research activity "Identifying sediment deposits from temperature signals," which falls under the Co-UDlabs project's Joint Research Activity "JRA 3- Improving Resilience and Sustainability in Urban Drainage solutions." The Zenodo dataset includes a ZIP file containing several CSV files documenting experiment data per Cycle and Pulse, sensor calibration data, and the standard methods employed.

The following steps were done in Renku to create a reproducible project:

1. Creating a Python environment setup
2. Importing data from external platforms
3. Setting up requirements
4. Wrangling and cleaning data
5. Jupyter Notebook and Python to illustrate data pipelines
6. Tracking data provenance with workflows
7. Sharing the project

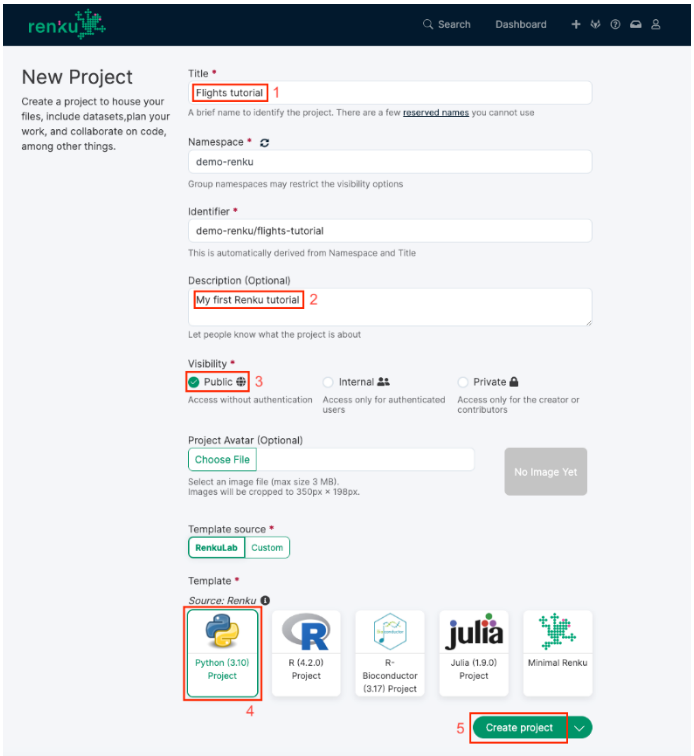

Figure 3: Screenshot of Renku's user interface pre-built environment options in Python, R, or Julia, along with varying visibility settings. (Source: Chavarria, 2023)

To achieve a fully reproducible workflow, we started the process using Renku web interface to setup a Python project. Renku offers a pre-built environment with a basic folder structure in Python, R, or Julia, and visibility options ranging from public, internal, to private (Figure 3) (1). The next step involved importing the dataset from Zenodo using the Renku CLI commands (2). With the data importing process, we identified and fixed errors in one of the CSV files uploaded in Zenodo repository (3).

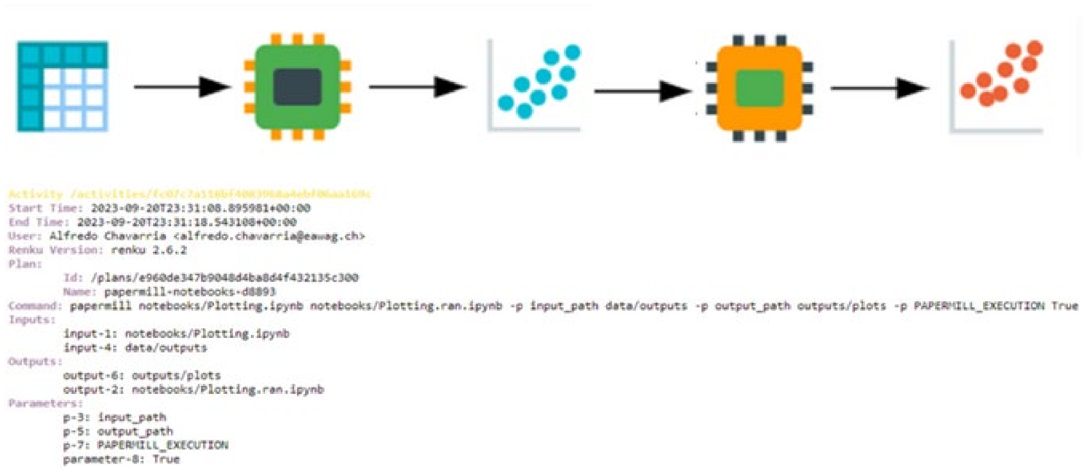

Figure 4: Workflow diagram and logs describing how the data cleaning of the Temperature Dataset serves as input for a subsequent step. (Source: Adapted from Renku Documentation, 2023)

To illustrate the data pipeline, we used Jupyter Notebook, which is integrated into RenkuLab's interface (4). Every code segment that processes input data and produces output data constitutes a workflow step in Renku. For the use case of the Temperature Dataset, we use the results of the data cleaning process as input data of a subsequent step as shown in Figure 4. These workflows can be labelled and listed in a ".yml" or ".yaml" files and run when needed (5). Finally, each Renku project created is mirrored as a GitLab repository. For sharing content, a project can be cloned via GitLab, enabling users to explore locally. A Docker file is also part of the options a project can be shared (6).

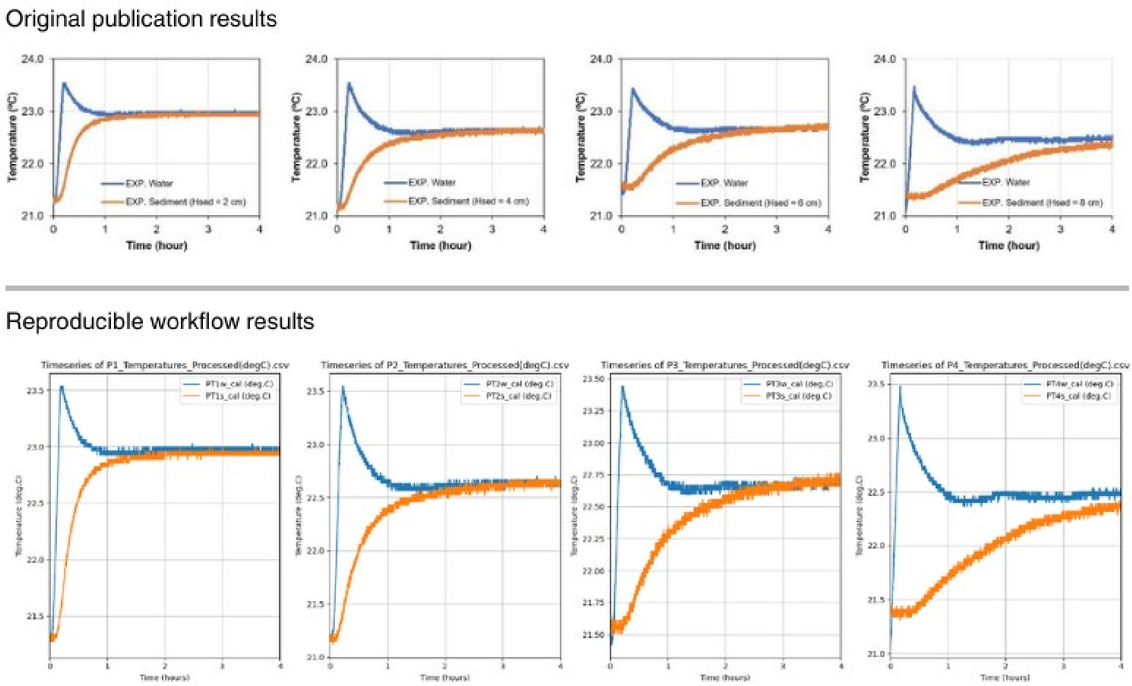

Figure 5: Comparison of results from the original publication and the reproducible workflow in Renku (Source: Chavarría, 2023)

As shown in the Figure 5, the results from the original publication can be compared to the ones obtained through the reproducible workflow. Reproducibility reinforces the reliability of the findings. Allowing other users to reproduce the results provides another layer of validation, minimizing the propagation of errors. As in our case, we identified an error in the structure of one of the CSV files, catching errors or inconsistencies improves the integrity of the results.

4.2. OGC Standards

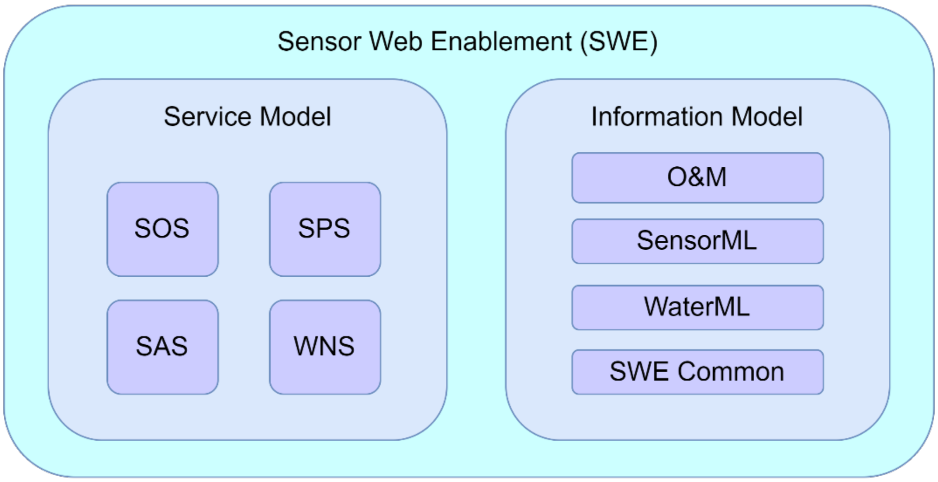

Figure 6: Sensor Web Enablement Framework (adapted from Jirka et al, 2009)

The UDS community have need of advancements in data harmonization. Achieving semantic interoperability is a crucial step in data harmonization, this ensures that users and implementers of different systems can grasp the semantics of the information provided by diverse systems. The Open Geospatial Consortium (OGC) has been developing location information standards to accomplish harmonization. It's noteworthy that geographic information is defined as "information concerning phenomena implicitly or explicitly associated with a location relative to the Earth" (Taylor et al., 2010), and hydrological phenomena fit into this category.

Under observational data, a lot of hydrological measurements can be covered. The OGC has developed a standards framework for observational data coming from sensors: the Sensor Web Enablement (SWE). SWE is a suite of standards designed to enable the discovery, access, and use of sensor observations, sensor descriptions, and digital representations of observed features in an interoperable and web-enabled manner. The SWE overarching framework divides its standards into two primary categories: the Service Model and the Information Model (Figure 6). The Information Model defines the data structures and relationships for representing sensor-related information, while the Service Model specifies interfaces, operations, and bindings for services that use the data structures from the Information Model.

The Information Model SensorML provides a framework to describe sensors and their processes. This includes details about how the sensor works, what it measures, its capabilities, and even its physical location. The Observation & Measurements (O&M) framework is an XML implementation of schemas for observations, and for features involved in sampling when making observations. An important part of observational data is placed on the procedure used to obtain the data and the resulting uncertainties, and the O&M standard addresses these points. Defined by O&M, an observation is "an action whose result is an estimate of the value of some property of the feature-of-interest, obtained using a specified procedure" (S. Cox, 2011). Adapting this, one can describe hydrological observations, for instance, turbidity in a river using a Nephelometer.

The hydrology community made progress in harmonizing information models through the WaterML standard. This standardized approach is based on the information model of the O&M standard and enhances data interoperability and facilitates seamless data exchange in hydrological applications. As a part of the WaterML2.0 standard, OGC has introduced four distinct document parts, each focused on a unique area:

- OGC® WaterML 2.0: Part 1- Timeseries
- OGC WaterML2.0: Part 2 – Ratings, Gaugings and Sections
- OGC® WaterML 2: Part 3 – Surface Hydrology Features (HY_Features) – Conceptual Model
- OGC WaterML 2: Part 4 – GroundWaterML 2 (GWML2)

And a Best Practices document for Water Quality:

- WaterML-WQ – an O&M and WaterML 2.0 profile for water quality data

The latter provides guidelines on how to configure XML documents for various water quality measurements, using O&M and WML 2 standards.

The XML structure described in Figure 7 represents an example of a water quality measurement (xml line 9) (S. J. D. Cox & Simons, 2014). The observation was done in 2005 (xml line 14), the Turbidity (xml line 27) was measured using a Nephelometer (xml line 19) sensor in the Yarragadee Aquifer (xml line 30), with a value of 100 NTU (xml line 34) at a water temperature (xml line 22) of 22.3°C (xml line 23).

```
1  <om:OM_Observation
2      gml:id="WQ_Measurement_Test01"
3      xmlns:om="http://www.opengis.net/om/2.0"
4      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
5      xmlns:xlink="http://www.w3.org/1999/xlink"
6      xmlns:gml="http://www.opengis.net/gml/3.2"
7      xsi:schemaLocation="http://www.opengis.net/om/2.0 http://schemas.opengis.net/om/2.0/observation.xsd">
8      <gml:description>Measurement test instance: turbidity measurement</gml:description>
9      <gml:name>Water quality measurement test 01</gml:name>
10     <om:type xlink:href="http://www.opengis.net/def/observationType/OGC-OM/2.0/OM_Measurement"/>
11     <om:phenomenonTime>
12         <gml:TimeInstant
13             gml:id="ot1t">
14             <gml:timePosition>2005-01-11T16:22:25.00+10:00</gml:timePosition>
15         </gml:TimeInstant>
16     </om:phenomenonTime>
17     <om:resultTime xlink:href="#ot1t"/>
18     <om:procedure
19         xlink:href="http://en.wikipedia.org/wiki/Nephelometer"/>
20     <om:parameter>
21         <om:NamedValue>
22             <om:name xlink:href=" http://environment.data.gov.au/def/property/water_temperature " xlink:title="water temperature"/>
23             <om:value xsi:type="gml:MeasureType" uom="http://qudt.org/vocab/unit#DegreeCelsius">22.3</om:value>
24         </om:NamedValue>
25     </om:parameter>
26     <om:observedProperty
27         xlink:href=" http://environment.data.gov.au/def/property/turbidity-water " xlink:title="water turbidity"/>
28     <om:featureOfInterest
29         xlink:href="http://environment.data.gov.au/groundwater/aquifers/yarragadee-aquifer"
30         xlink:title="Yarragadee Aquifer"
31         xlink:role="http://environment.data.gov.au/def/featuretype/HydrologicUnit"/>
32     <om:result
33         xsi:type="gml:MeasureType"
34                                 uom="http://qudt.org/vocab/unit#NephelometricTurbidityUnit">100</om:result>
35 </om:OM_Observation>
```

Figure 7: XML structure using the O&M standard highlighting a water quality measurement (Source: Chavarría, 2023)

To align the UDS community's practices with the OGC standards, the UDS data must comply with the O&M, WaterML, and SensorML requirements set by OGC. This comprises:

1. Conversion of Sensor Data Formats: The current and most used method of data representation and exchange, especially for sensor measurements, is through CSV files. In compliance with OGC standards, we need to transition from CSV representations to XML-based structures. This will require converting existing CSV files and adjusting the data logging methods to output XML files that align with the O&M and WaterML requirements.
2. Metadata and Additional Information: To adopt the OGC standards, the Co-UDlabs members have the required degree of expertise to understand and input specific data mandated by the OGC, such as detailed metadata, this includes but is not limited to sensor names, reference systems, observation times, variables names and measurement units. Ensuring that UDS data includes this metadata also increases the quality, traceability, and usability of our datasets.

While OGC standards have made important advances in harmonizing information models for hydrological observations, not every UDS measurement aligns with existing OGC standards. Looking ahead, extending WaterML to encompass urban drainage observations emerges as a potential avenue for the UDS community. However, active community participation is key. A good starting point is gathering datasets and field knowledge that describes the vocabulary, data structure, and measurement systems specific to urban drainage systems. The Co-UDlabs' working packages, including the Transnational Activities (TA) and Joint Research Activities (JRA), provide valuable test cases.

4.3. Integration of UDMT toolbox

Co-UDlabs WP6 developed the UDMT (Urban Drainage Metrology Toolbox), aiming to facilitate the adoption and use of best practices in metrology applied in the field of urban drainage. It is indeed observed that practices to use sensors and to process data in this field are far from being harmonized and shared, which adds difficulties and blockages in sharing using and reusing data sets.

The UDMT includes five blocks of functions:

1. Sensor calibration / Correlation: this block provides various methods to determine i) calibration functions (based on a data set of outputs of a sensor submitted to standards or certified values), and ii) correlation functions (based on a data set of values given by a sensor and corresponding values obtained e.g., with laboratory analyses of samples).
2. Calibration / Correlation correction: this block allows to convert raw values provided by a sensor into corrected values according to previously determined calibration or correlation functions. In addition, uncertainties in corrected values are estimated.
3. Uncertainty assessment: this block allows to apply standard methods (type A, type B, Monte Carlo) to various data sets. In addition, the variograph method is proposed to estimate uncertainties in integrated values (e.g., sums, means, etc.).
4. Data validation: this block provides a set of automated tests to help the user to validate data according to various criteria.
5. Tracing experiments: this block allows to calculate a discharge from experimental data collected during tracing experiments. Tracing experiments are useful to qualify flowmeters in urban drainage systems.

The UDMT was initially developed as a free access webapp, based on Matlab.[1]

In addition to the online version, the UDMT is also available as executable applications that have been compiled for both Microsoft Windows and Apple Mac operating systems.[2]

In the next phase of WP2, the integration of the UDMT in the WP2 Guidelines for harmonization will be investigated. If sensors are calibrated with standards protocols and their data processed according to standard methods, if harmonized methods are applied to correct raw data delivered by sensors by means of standardized calibration and correlation functions, if uncertainty assessment is systematically applied with standard methods, and if data are submitted to validation with standards tests and criteria, it will significantly help to generate data sets complying with the FAIR principles.

A first step toward the integration of the UDMT in WP2 will be the application of data formats and file formats developed in WP2 (see Section 3.2).

5. Expected Outcome and Contribution to the Objectives of Co-UDlabs

We expect that the main outcomes of the T2.1., the definition of FAIR standards, common protocols and methods for data harmonization and re-use, are fundamental for our endeavours to organize an effective data management system for the data collected during the project. Also, they are vital to standardize experimentation and ensure consistent high-quality data collection.

First, selecting Zenodo as a data sharing platform is the backbone to ensure seamless data sharing within the Co-UDlabs community. While this is not (yet) a community-driven data sharing platform, such as materialscloud.org, it has many advantages which directly contribute to building a collaborative European Urban Drainage innovation community, especially regarding the open availability and re-use of data. Relevant advantages are the aspect of long-term data preservation, which ensures that Co-UDlabs data remains accessible and intact over time, making The Co-UDlabs community on Zenodo reliable source for archiving and sharing urban drainage data. Also, Zenodo supports a variety of metadata standards and data formats, making it easy to integrate with other repositories, data management tools, and research platforms. Last, but not least, the platform allows multiple contributors to collaborate on a community, making it a useful platform for research teams or collaborations.

Second, unifying terminology and naming conventions, as well as a close integration with the European Open Science Cloud (EOSC), and data licensing, are critical aspects of standardizing data practices. Addressing these issues will significantly contribute to achieving a collaborative urban drainage community. We expect that we can, at least partly, standardize the terminology and naming conventions of monitoring data for important system variables, such as water levels, flows, and basic water quality variables, which will ensure that researchers will use common language and keywords when describing their data. This will make it easier for others to find and interpret datasets, especially in online data portals. In our view, integrating the Co-UDlabs data management practices with the European Open Science Cloud (EOSC) framework holds great promise to create a unified approach to data sharing

[1] The webapp of the UDMT is available online on Co-UDlabs' website: https://co-udlabs.eu/research/tools-and-outputs/.
[2] The exe versions work similarly to the online version. They can be downloaded here: https://u.pcloud.link/publink/show?code=kZegnQVZM53qyjRJ4cHn7Pi5WzrR9HJ0PL4V, and later installed on personal computers by the user. A detailed user manual with examples is also available at:https://u.pcloud.link/publink/show?code=kZegnQVZM53qyjRJ4cHn7Pi5WzrR9HJ0PL4V.

and archiving in the urban drainage community. If we can indeed successfully create an urban drainage data sharing community, possibly under the umbrella of established organizations, such as the International Working group of data and models (IWGDM) or the IWA/IAHR Joint Committee of Urban Drainage, researchers across Europe, and possibly globally, can align their practices with the EOSC, contributing to effective re-use of data and effective knowledge generation. This would be a paradigm shift in comparison to today's culture of ad-hoc collaboration.

Third, on a technical level, we expect that the Renku platform will be an invaluable asset in enhancing the quality and efficiency in how the urban drainage community deals with research data. We expect that we can deliver relevant use cases, e.g., rainfall data processing or assessing the usability of an open dataset for flood protection, which demonstrate the advantages of i) using git for efficient version control, allowing multiple collaborators to work simultaneously and maintain data and code consistency, ii) having data versioning features to document specific changes to datasets, iii) collaborative Jupyter Notebooks and RStudio environments to foster teamwork and knowledge sharing, and iv) documenting the provenance of derived datasets to promotes transparency and trust in our collaborative research initiatives.

On a governance level, we expect that the clear recommendation for a CC-C0 licence (license waiver) will ensure that data in the urban drainage community are shared with appropriate usage permissions and restrictions, contributing to the "Accessible" and "Reusable" aspects of the FAIR principles.

Regarding the contribution to overall Co-UDlabs goal and objective, the expected results from task 2.1. will directly support the objectives of WP2. Specifically, they will "standardize experimentation and operation of the research infrastructures and ensure consistent high-quality data collection [...] during the Co-UDlabs project." (O2.1), create "an effective data management system for the data collected during the project. This system will include curation, preservation and provision of access to the collected data and metadata as well as the incorporation of new standards and data protocols developed in the JRA actions." (O2.2) and enable the partners and the urban drainage community to "exchange best practices and know-how among the project staff and the participants working in Research Infrastructures." (O2.3).

All these specific outcomes will directly contribute to fulfilling the Co-UDlabs goals, internally-oriented activities to ensure interoperability such as harmonization and an overarching data management, as well as externally-oriented activities to foster smart governance and policy as well as public access to data.

On a higher level, they will mostly contribute to fostering "a culture of co-operation between RIs and the urban drainage community through a set of coordinated Networking Activities [...], which help to develop a more inclusive, open and efficient research and innovation environment." (Main Objective 1). In addition, they will partly contribute to the Main Objective 3 "to enlarge and strengthen the quality and quantity of the services offered at European level by Co-UDlabs by [...] digital water data analysis tools".

As a result, the anticipated outcomes hold significant socio-environmental implications and are in harmony with broader EU policies. This is underscored by the recognition of technologies and the concept of 'hybrid grey and green infrastructure' as key components for realizing a Water Smart Society, as outlined in the Strategic Innovation and Research Agenda (SIRA) of the Water Europe platform. Truly FAIR data enhances the goal of the European Open Science Cloud for seamless, standardized data access and sharing, which directly advances EU research aims.

In doing so, these efforts align seamlessly with the overarching vision of Horizon Europe, which seeks to foster a sustainable, equitable, and prosperous future, in line with European values, by addressing climate change, contributing to the achievement of Sustainable Development Goals, and bolstering the Union's competitiveness and growth. This process aims at having the following impact on current research:

- Accelerating Scientific Discoveries: Interoperable data empowers researchers to combine and analyze a wide array of datasets, expediting scientific breakthroughs through in-depth analysis.
- Cost Efficiency: The reusability of data curtails the need for redundant data collection, resulting in resource savings that harmonize with the EU's drive for efficient research practices.
- Multidisciplinary Support: Fair data practices foster effective collaboration across diverse fields, aligning perfectly with the EU's emphasis on multidisciplinary research objectives.
- EOSC Integration: Fair data plays a pivotal role in enhancing the European Open Science Cloud's vision of creating a seamless and standardized environment for data access and sharing, thereby propelling EU research goals.
- Open Science Alignment: Fair data perfectly complements open science principles, promoting transparency, sharing, and accessibility within the realm of EU research.

6. APPENDIX

6.1. Licences for open research data

In Table 5, it is shown that the available open datasets have very different licenses which regulate how the data can be re-used. This is important, because the status of research data and its eligibility for copyright protection can have significant implications for how it can be shared and published. The fundamental question is whether the research data set qualifies for copyright protection. If it does, then you have the option to publish it under a Creative Commons (CC) license. Conversely, if your data does not meet the criteria for copyright protection, you cannot apply a CC License to it.

6.1.1. Conditions for protecting research data by copyright

Several conditions must be met for research data to be considered protected by copyright:

Firstly, the data must be created or collected by a human being, such as through interviews, observations, or surveys. Data generated by machines or animals typically does not fall under copyright protection.

Secondly, the data must be perceivable to the senses, meaning it should be readable, listenable, or touchable. In most research data cases, this requirement is satisfied as the data can be considered as expressed.

Lastly, originality is a critical criterion. While evaluating originality can be challenging, it is essential. For example, data gathered through a survey may not be inherently original, but the way it is organized, such as in an Excel spreadsheet, might possess original elements. Similarly, the organization and presentation of data can contribute to its originality, making it eligible for copyright protection.

From our consortium's perspective, monitoring data typically does not originate from an intellectual process akin to art or music creation. Consequently, they often do not qualify for copyright protection under most circumstances. There may be exceptions, such as data collections that involve curation and intellectual input. Even in these instances, applying a Creative Commons (CC) license to data can be somewhat unwieldy, as it may reflect a lack of understanding on the part of the licensor. Moreover, it introduces complexities when sharing multiple datasets, requiring constant vigilance to ensure no license violations occur.

As a solution, we strongly recommend that all Co-UDlabs data be placed in the public domain, effectively waiving copyright protection. This aligns with the principles of a standard CC0 license, which signifies that the data is freely available for public use without any encumbrances related to copyright.

6.1.2. Identifying examples of data protected by copyright with the DMLawTool

A practical resource for identifying examples of data protected by copyright can be found at the DMLawTool, which mainly targets researchers in the humanities and social sciences. The DMLawTool has been developed as an interactive decision-tree as part of the swissuniversities P-5 programme "Scientific Information." Researchers can use this tool to navigate essential legal considerations associated with data management. Upon completion of the guidance process, the tool offers various solution pathways to determine how researchers can handle their research data and potentially archive them in a repository.

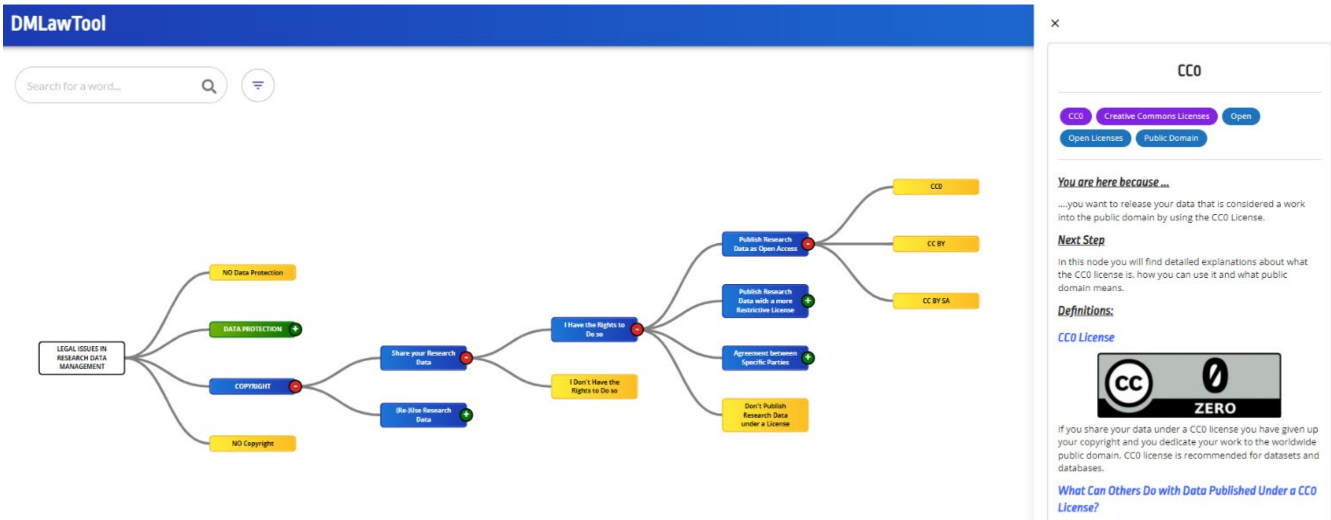

Figure 8: DMLawTool (Source: Chavarría, 2023)

The DMLawTool, an interactive decision-tree resource, helps researchers, primarily in the fields of humanities and social sciences, to navigate essential legal considerations associated with data management. It has been developed through a collaboration between the Università della Svizzera italiana and the University of Neuchâtel, as part of the swissuniversities P-5 programme "Scientific Information." It is available as open-source software under the CC-BY-SA 4.0 license.

The authors provide an important disclaimer (*DMLawTool*, 2023):

> *"This tool is a basic guide with the only purpose of giving a general understanding on the main legal aspects of Copyright and Data Protection according to Swiss legislations and to the European General Data Protection Regulation as of March 2021.*
>
> *The tool is developed in a practical manner, and it is written in a non-legal language. Its purpose is not to provide complete and tailored legal advice to the user's case, but to help researchers identify legal issues, and highlight clues that users may consider or implement with the help of specialists when necessary.*
>
> *If not indicated otherwise, the examples used are invented and have the only purpose of a better illustration of the information provided. All the information provided in this tool does not, and is not intended to, constitute legal or other professional advice.*
>
> *The user of this tool acknowledges that each situation must be judged on a case-by-case basis and must seek legal advice from a competent attorney to resolve their specific case with respect to any particular legal matter."*

Understanding these details concerning the copyright status of research data is vital for researchers when deciding how to license and share their valuable work while ensuring compliance with legal regulations.

7. References

Anta, J., Ciambra, A., Tait, S., Moreno, A., Rieckermann, J., Brüggemann, T., Bertrand-Krajewski, J.-L., Ellerbaek, J., Brelot, E., & Hayaoui, M. (2022). *Co-UDlabs—D10.1 Periodic Technical Report (M18) Part B* [1st periodic report].

Anta, J., Regueiro-Picallo, M., Naves, A., & Pernas, R. (2022). *Co-UDlabs_WP8_T812_UDC_001. Identifying sediment deposits from temperature signals* [dataset]. https://doi.org/10.5281/zenodo.7827667

ATV. (1999). *ATV-M 143 -2: Optische Inspektion – Inspektion, Instandsetzung, Sanierung und Erneuerung von Abwasserkanälen und -leitungen Teil 2: Optische Inspektion*.

Biere, Y. (2013). *Vergleichende Bewertung von Abweichungen in Kanalinspektionsergebnissen aus der Erst- und Wiederholungsinspektion (Comparative assessment of deviations in sewer inspection results from initial and repeat inspections). Bachelor thesis*. Ruhr-Universität Bochum.

Brandl, L. & Beschreibung. (2017). *Modellierung und Prognose der Zustandsveränderungen von Bauteilen der Abwasserkanalisation (Description, modelling and prognosis of changes in the condition of components of the wastewater sewer system* [Master thesis.]. Ruhr-Universität Bochum.

Caradot, N., Sonnenberg, H., Riechel, M., Matzinger, A., & Rouault, P. (2013). The influence of local calibration on the quality of UV-VIS spectrometer measurements in urban stormwater monitoring. *Water Practice and Technology*, *8*(3–4), 417–424. https://doi.org/10.2166/wpt.2013.042

Cox, S. (2011). *Observations and Measurements—XML Implementation. Version 2.0.* [Report]. Open Geospatial Consortium. https://doi.org/10.25607/OBP-627

Cox, S. J. D., & Simons, B. A. (2014, December 2). *WaterML-WQ – an O&M and WaterML 2.0 profile for water quality data* [Best Practice]. Open Geospatial Consortium. https://docs.ogc.org/bp/14-003/14-003.html

Dirksen, J., Clemens, F. H. L. R., Korving, H., Cherqui, F., Le Gauffre, P., Ertl, T., Plihal, H., Müller, K., & Snaterse, C. T. M. (2011). The concistency of viusal sewer inspection data. *Structure and Infrastructure Engineering*, 1-15. https://doi.org/10.1080/15732479.2010.541265

Dotto, C. B. S., Mannina, G., Kleidorfer, M., Vezzaro, L., Henrichs, M., McCarthy, D. T., Freni, G., Rauch, W., & Deletic, A. (2012). Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling. *Water Research*, *46*(8), 2545–2558. https://doi.org/10.1016/j.watres.2012.02.009

DMLawTool. (2023). https://dmlawtool.ccdigitallaw.ch/disclaimer

DWA. (2005). *DWA-M 159—Kriterien zur Materialauswahl für Abwasserleitungen und -kanäle, Merkblatt*.

DWA. (2010). *DWA-M 149-5 Zustandserfassung und -beurteilung von Entwässerungssystemen außerhalb von Gebäuden. Teil 5: Optische Inspektion. Deutsche Vereinigung für Wasserwirtschaft*.

DWA. (2013). *DWA-M 149-2 Zustandserfassung und -beurteilung von Entwässerungssystemen außerhalb von Gebäuden. Teil 2: Kodiersystem für die optische Inspektion. Deutsche Vereinigung für Wasserwirtschaft, Abwasser*.

DWA. (2014). *DIN EN 13508-2 "Untersuchung und Beurteilung von Entwässerungssystemen ausserhalb von Gebäuden—Teil 2: Kodiersystem für die optische Inspektion."* https://shop.sia.ch/europ%c3%a4ische%20normen/ingenieur/sia%20190.232+a1/d/D/Product

Flood Forecasting Centre. (2023, January 25). GOV.UK. https://www.gov.uk/government/organisations/flood-forecasting-centre

Fuchs, L., & Harms, R. W. (1986). HYSTEM-EXTRAN - Model Application in Scientific Investigations. In C. Maksimovic & M. Radojkovic (Eds.), *Urban Drainage Modelling, Proc. Of the International Symposium on Comparision of Urban Drainage Models with Real Catchment Data*. Pergamon Press.

Gee, B. (2023, May 30). *Good data practices: Removing barriers to data reuse with CC0 licensing*. Dryad News. https://blog.datadryad.org/2023/05/30/good-data-practices-removing-barriers-to-data-reuse-with-cc0-licensing/

Gutiérrez, J. M., Jones, R. G., Narisma, G. T., Alves, L. M., Amjad, M., Gorodetskaya, I. V., Grose, M., Klutse, N. A. B., Krakovska, S., Li, J., Martínez-Castro, D., Mearns, L. O., Mernild, S. H., Ngo-Duc, T., van den Hurk, B., & Yoon, J.-H. (2021). *Atlas* [Repository]. Interactive Atlas of the Intergovernmental Panel on Climate Change. https://interactive-atlas.ipcc.ch/

Hochrheinbrücke. (2023). In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Hochrheinbr%C3%BCcke&oldid=229760636

Lepot, M., Torres, A., Hofer, T., Caradot, N., Gruber, G., Aubin, J.-B., & Bertrand-Krajewski, J.-L. (2016). Calibration of UV/Vis spectrophotometers: A review and comparison of different methods to estimate TSS and total and dissolved COD concentrations in sewers, WWTPs and rivers. *Water Research*, *101*, 519–534. https://doi.org/10.1016/j.watres.2016.05.070

Mars Climate Orbiter. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Mars_Climate_Orbiter&oldid=1181582662

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 System Science Data*, *13*(10), 4779–4798. https://doi.org/10.5194/essd-13-4779-2021

NEN 3399:2015 nl. (2015). https://www.nen.nl/nen-3399-2015-nl-197752

Nunnenmacher, L. (2023). *Personal Communication*.

Prodanović, Dušan, & Branisavljević, Nemanja. (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.

Radojkovic, M., & Maksimovic, C. (1986). *Urban Drainage Catchments: Selected Worldwide Rainfall-Runoff Data from Experimental Catchments* (First Edition). Pergamon Pr.

Roskar, R., Ramakrishnan, C., & Volpi, M. (2023). *Renku: A platform for sustainable data science*. NeurIPS | 2023 Thirty-seventh Conference on Neural Information Processing Systems.

Tacconelli, E., Gorska, A., Carrara, E., Davis, R. J., Bonten, M., Friedrich, A. W., Glasner, C., Goossens, H., Hasenauer, J., Abad, J. M. H., Peñalvo, J. L., Sanchez-Niubo, A., Sialm, A., Scipione, G., Soriano, G., Yazdanpanah, Y., Vorstenbosch, E., & Jaenisch, T. (2022). Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response. *The Lancet Regional Health – Europe*, *21*. https://doi.org/10.1016/j.lanepe.2022.100467

Taylor, P., Walker, G., Valentine, D., & Cox, S. (2010, July 14). *WaterML2.0: Harmonising Standards for Water Observations Data*. JRC Publications Repository. https://publications.jrc.ec.europa.eu/repository/handle/JRC56671

von Waldow, H. (2017, March 13). There is no research data for archiving. *Ckan4rdm*. https://ckan4rdm.wordpress.com/2017/03/13/there-is-no-research-data-for-archiving/

Williams, M., Bagwell, J., & Nahm Zozus, M. (2017). Data management plans: The missing perspective. *Journal of Biomedical Informatics*, *71*, 130–142. https://doi.org/10.1016/j.jbi.2017.05.004

WRc. (2013). *Manual of Sewer Condition Classification, MSCC* (5th ed.). htts://wrcknowledgestore.co.uk