

**Interreg
Danube Region**



**Co-funded by
the European Union**



Output 1.3

Testing and demonstration of a hazardous substances database as operative transnational tool to perform regular transnational modelling-based pollution assessment and to support emissions models and risk assessment

Pilot action jointly developed and implemented by:

TU Wien (AT), Budapest University of Technology and Economics (HU), International Commission for the Protection of the Danube River, Bulgarian Water Association (BG), National Administration „Romanian Waters” (RO), Water Research Institute (SK), Jozef Stefan Institute (SI), Center for Ecotoxicological Research Podgorica (ME), Ukrainian Hydrometeorological Institute State Service on Emergencies and National Academy of Science (UA), Croatian Waters (HR), Public Institution “Vode Srpske” (BA), Environment Agency Austria (AT), Jaroslav Černi Water Institute (RS)

Authors:

S. Kittlaus, S. Osten, A. Kovacs, M. Kardos, A. Clement, T. Lajkó, O. Zoboli

December 2025

Table of content

| | |
|--|----|
| Abstract | 1 |
| 1 Introduction..... | 2 |
| 1.1 Hazardous substance management in the Danube River Basin | 2 |
| 1.2 The Danube Hazard m ³ c concentration database | 2 |
| 1.3 Requirements for an improved database and pollution evaluation tool | 3 |
| 2 The joint development process..... | 4 |
| 2.1 Kick-off and collection of pre-existing national data structures..... | 4 |
| 2.2 Redesign of the database..... | 4 |
| 2.3 Testing of the database..... | 5 |
| 2.4 Development of a stewardship for the database | 5 |
| 2.5 Operation of the database..... | 6 |
| 2.6 Use of the database for the validation of transnational model results..... | 6 |
| 2.7 Design of a pollution evaluation tool (PET)..... | 8 |
| 3 Lessons learned and outlook..... | 10 |

Abstract

Within the previous INTERREG Danube project *Danube Hazard m³c*, a large PostgreSQL database was developed to collect and harmonize hazardous substance concentration data from multiple environmental compartments across the Danube River Basin. The database comprised 10.7 million measurements from over 383,000 samples at approximately 25,000 sampling sites and proved valuable for assessing data availability, identifying monitoring priorities, and supporting emission modelling. However, post-project evaluation revealed limitations in usability, documentation, and consistency of controlled vocabularies. Additional challenges arose from aggregated data handling, particularly in the presence of values below the limit of quantification, which restricts reuse beyond original purposes.

The Tethys project thus carried out a pilot action, in which the partners jointly redesigned, expanded and improved the existing database to make it a more widely applicable instrument. Moreover, additional tools and workflows were designed and integrated to support among others the import of large datasets, thorough quality control assurance and the visualization, processing and evaluation of the data.

The early involvement of and close coordination with the ICPDR experts ensured that the database is in line with the ICPDR needs, compatible with existing ICPDR data management tools and fully transparent and accessible for the ICPDR experts and stakeholders. The ICPDR recognizes its high value and plans to explore concrete ways to adopt it as its own potential future operational tool to support cycling emission modelling, establishment of emission inventories and other analyses of water contamination at the basin-wide scale.

1 Introduction

1.1 Hazardous substance management in the Danube River Basin

Despite the substantial progress achieved in controlling hazardous substances (HS) pollution in the Danube River Basin, further efforts are needed in the future to sustainably manage the problem. Narrowing the information gap related to HS pollution is a key aspect, the state-of-the-art knowledge regarding monitoring and chemical emissions needs to be massively improved. Moreover, the water management sector must adapt its policies and measures to reflect to the latest EU requirements and to establish an enabling regulatory framework that can support and control the implementation of these measures. In addition, in some Danube countries, substantial lack of institutional capacity, unclear responsibilities and insufficient intersectoral dialogue hinder the establishment of efficient management. Furthermore, the pressure of climate change impacts on water quality management is rapidly growing, both high intensity rainfall events and drought periods may cause adverse effects if they are not counteracted by adaptation measures.

For effective control of HS water pollution, policies shall take actions in three fundamental areas, all representing an essential part of managing HS pollution. They should first establish an appropriate monitoring system to identify the problem, including a well-designed and harmonized monitoring program, modern equipment for measuring and a **well-structured open access database of HS concentrations**. Secondly, they need to develop a comprehensive emission inventory building on the database to determine the main pollution sources and pathways and to quantify the pollution fluxes within the catchments. Finally, they must put in place appropriate control measures to minimize pollution, targeting the main sources and pathways.

1.2 The Danube Hazard m³c concentration database

Within the INTERREG Danube project “Danube Hazard m³c” a database for collection of hazardous substance concentrations was developed to support emission modelling. Data from multiple environmental compartments in many countries within the Danube river basin were collected, harmonized and stored in a PostgreSQL database. These were on the one hand aggregated concentration data and on the other hand single values originating from national monitoring, other research projects and from the monitoring conducted within that project itself. A detailed description can be found in Kittlaus et al. (2024)¹ and the database itself including all openly licensed data is available for download in form of a PostgreSQL-dump: <https://doi.org/10.48436/kxxnm-xss93>

The results achieved were remarkable, with the database containing 10.7 million concentration measurements in over 383 000 samples from approximately 25 000 sampling sites from the Danubian countries. This database proved very valuable to assess, quantify and compare data availability in the Danube basin, to identify priorities for future monitoring programmes and it also allowed to derive essential input and validation data for emission modelling. Nevertheless, the fact that most of the resources and efforts within the Danube Hazard m³c project had to be invested in the search, collection, harmonisation and evaluation of the data, after the project completion a critical evaluation of the technical tool itself revealed that some aspects were suboptimal for its sustainable and efficient use. The interfaces were designed for expert usage and some naming and controlled vocabularies

¹ Kittlaus, S., Kardos, M. K., Dudás, K. M., Weber, N., Clement, A., Petkova, S., Sukovic, D., Kučić Grgić, D., Kovacs, A., Kocman, D., Moldovan, C., Kirchner, M., Gabriel, O., Krampe, J., Zessner, M. and Zoboli, O. 2024 A harmonized Danube basin-wide multi-compartment concentration database to support inventories of micropollutant emissions to surface waters. *Environ Sci Eur*, **36**(1). <https://doi.org/10.1186/s12302-024-00862-4>

exhibited traits of the development process, meaning not having harmonized naming of columns and slightly different data structures for similar types of data.

Furthermore, the technical documentation of database presented improvement potential and, it became apparent that a more accurate definition and documentation of the controlled vocabulary was necessary, as there were instances in which different understanding of concepts and terminology among project team members led to diverging naming of similar things. This can lead in turn to a loss of information or to errors in data analysis and processing.

Moreover, in the process of data evaluation the collected aggregated data gave place to multiple issues and challenges. A particular challenge concerned the treatment of values below the analytical limit of quantification (LOQ) during data aggregation. For data received in aggregated form, different approaches applied during data aggregation hinders a meaningful disaggregation and a use of these data aside from their original purpose.

1.3 Requirements for an improved database and pollution evaluation tool

The Tethys project provided the opportunity to carefully redesign and extend the database developed in the Danube Hazard m³c project to make it a more widely applicable instrument. Additional tools for supporting the work with the database and the use of the data contained in it were considered beneficial, to curate and to leverage the full benefits of such an extensive data collection.

The following specific requirements were identified as relevant for the further development process:

- Use of free/open source software
- Better user interfaces
- Better documentation
- Harmonization with EU WISE and UWWTD reporting data formats and controlled vocabularies to reduce effort for data preparation.
- Use of best practises (controlled vocabulary, ISO standards for item registration)
- Clean separation of compartment specific data contents via views, but coherent data storage
- Quality assurance workflow
- Clear licensing of datasets to allow reuse where possible
- Code and software versioning to maintain a clear overview of development steps
- Setup of a development process which allows creation and update of databases on servers not directly accessible for the development team
- Common development of the databases for national use and the transnational database to keep data structure similar and allow for swift data exchange
- Better data management starting at import through the introduction of unique, foreign and not null constraints.
- Helper views to find controlled vocabularies or table and column definitions

2 The joint development process

2.1 Kick-off and collection of pre-existing national data structures

A first meeting on the development of the transnational database took place in early February 2024 between the ICPDR and the lead partner TU Wien, where expectations of the ICPDR regarding the tools and possibilities of their adaption into the work of the ICPDR were discussed.

The common development process involving all partners was kicked off in February 2024 with an online meeting where the project team members of all project partners involved in activity 1.4 (transnational database) and A1.5 (national databases) met and the partners presented each other their expectations and wishes for improving the data management.

Following this meeting, the database schema of the ICPDR database of monitoring data, which was under reconstruction at this time and samples of national data formats were collected to assess which specificities might need to be considered and what meta data are mostly available.

2.2 Redesign of the database

A thorough redesign of the previously available database schema from the Danube Hazard m³c project was undertaken focusing on harmonised naming conventions and compatibility with other data formats used in the field. Some examples and aspects on this are given below:

- The determinants (mostly substances but partly also physicochemical parameters) are a controlled vocabulary in the database which can be extended by the users if needed. Here a wide variety of substance identifiers, starting from the well-known CAS registry number, over the EC number used by the European Chemicals Agency, different names (common ones as well as those according to IUPAC), abbreviations and finally keys used in cheminformatics can be added.
- The analytical methods used to derive the concentrations from environmental samples were based on the NORMAN network database system attributes, which allow to classify applied methods regarding their status of establishment and quality assurance procedures applied to allow for comparability of results.
- All data of the same type, e.g. concentration measurements in different environmental compartments and matrices are stored in one table to have identical data structures everywhere, except where differences are explicitly needed, e.g. different units of measurement for solid and liquid matrices.
- To keep the performance of the database (speed of delivering requested data) at a high level, even when there are millions of datasets introduced, the biggest tables were technically partitioned in the background according to environmental compartments, so e.g. surface water measurements are stored in a different table partition than groundwater measurements. Nevertheless, all constraints and checks are defined at the level of the overall table to ensure harmonised data structure and easy management.

As the project partnership had extensive expertise in laboratory analysis and data management, this know-how was integrated in the development process by targeted online meetings for clarifying questions regarding documentation of chemical analysis. JCWI presented their own advanced laboratory information system and shared information about the database structure behind it and experiences in working with the system.

As targeted adaptation to maximise synergies between the transnational database and national databases, the system was updated to allow for **bilingual controlled vocabularies**. This way the

national application of the database can work with translated controlled vocabulary in national language, while in the background the English terms stay available for data exchange with the transnational database without need for additional translation.

As the database is mainly developed to support emission modelling of hazardous substances, the partners responsible for emission modelling on the national and transnational scale were invited for a dedicated meeting for clarifying the way the database can optimally support the national and transnational modelling.

2.3 Testing of the database

In August 2024 a database prototype was presented by TU Wien to the project partners. This test database was installed on a TU Wien server and all involved project partners received access to it via VPN connections. The free open-source software DBeaver² was recommended as graphical user interface for the database.

A documentation on how the database was structured and how it can be used was made available in a wiki page on the TU Wien GitLab server. This documentation (including a FAQ), as well as the code and the ticket system used by the developers, were made available to allow all partners to follow the development.

To facilitate an easy start in using the database, nine video tutorials were recorded by the team of TU Wien and made available in the project cloud space. These tutorials covered the following topics:

1. Managing multiple database connections in DBeaver
2. Structure of the database
3. Data import strategy and workflow
4. Data import practise
5. Data import via the DBeaver import wizard
6. Data import via R
7. The quality assessment workflow – Introduction
8. The quality assessment workflow – perspective of the quality assessor
9. The quality assessment workflow – perspective of the importing user

The testing phase lasted three months, after which the final development stage was launched. The test database was kept available as a sandbox for the teams of the project partners to play around and test features of the tool without any risk of corrupting the actual final tool under development and the data it contained.

2.4 Development of a stewardship for the database

The transnational database is considered as a highly valuable tool for the ICPDR, providing useful information on concentrations of target hazardous substances in various environmental matrices and supporting the periodical assessments of basin-wide chemical emissions. Therefore, the ICPDR ensured close coordination with the project partners to support the development of the database in line with the needs of the ICPDR and the existing ICPDR data management tools and procedures.

The database was introduced to and discussed with relevant ICPDR experts (Secretariat and the Pollution Management Expert Group) at several technical ICPDR meetings to gain their feedback. The ICPDR experts expressed full support to the initiative of the future adoption and use of the database that shall be an essential operational tool for collecting, processing, analysing and disclosing data on

² DBeaver Community edition: <https://dbeaver.io>

chemicals' concentrations and for supporting periodic modelling activities on the basin-wide level. In line with its transboundary coordination role, the ICPDR is willing to facilitate future potential updates of the database, closely linked to the updates of the national databases to be carried out by the Danube countries. In addition, the ICPDR will consider the possibility of integrating certain parts of the transnational database into the ICPDR water quality database. Moreover, the ICPDR intends to support the dissemination of relevant data to the interested stakeholders through its existing platforms.

Moreover, the ICPDR Delegations were informed about the database and the ICPDR was requested to explore how to make use of the technical tools developed in Tethys for future basin-wide pollution management purposes. In line with this request, ICPDR will organize a special technical session with the project representatives in spring 2026 to discuss the details and necessary steps of the adoption and future use of the project results at the ICPDR level, including the database.

2.5 Operation of the database

While the development of a stewardship for the long-term use of the database was still under development, the first instance of the final database was set up at the server of TU Wien and access was granted to all partners to import data.

As the first major dataset, the monitoring data created within the Tethys project itself were imported. Each partner uploaded the meta data for monitoring locations and samples, while TU Wien imported the results of the chemical analyses from its own laboratory and from the other laboratories involved in the monitoring activity of the project.

In September 2025 a two-day hands-on workshop on data management and evaluation was organized at TU Wien for the project partners, which covered among others the following topics:

- Status of the transnational database
- Aspects of data quality assessment
- Hands-on session for data import
- Data evaluation using SQL and R
- What's behind the user interfaces: Main aspects of technical implementation

Aside from the Tethys project monitoring data, several partners imported further national datasets into the transnational database.

2.6 Use of the database for the validation of transnational model results

The validation of the transnational pathway-oriented MoRE model, implemented in the project for the whole Danube River Basin, was based on a combination of large-scale monitoring programmes and project-specific datasets covering both water quality and discharge information across the Basin.

Concentration data were derived from multiple sources, including the Danube Hazard m³c inventory, the Transnational Monitoring Network (TNMN), the Joint Danube Surveys (JDS3 and JDS4), and the Tethys project. Together, these datasets provided complementary spatial and temporal coverage for metals, pharmaceuticals, and PFAS in surface waters. The majority of these data sources are already included in the transnational database, while further datasets are being integrated as part of its ongoing development.

Two main validation configurations were applied. JDS-based validation relied on datasets representing low-flow summer conditions, while validation under mean discharge conditions combined monitoring data from TNMN, Danube Hazard m³c, and Tethys.

Discharge data used for load calculations were primarily obtained from the Interreg Danube Water Balance (DWB) project, providing measured daily flow data for a large number of stations. These data were complemented by model-based discharge estimates from the CWatM model and long-term discharges from the ICPDR–MONERIS modelling framework. The combined use of measured and modelled discharge datasets ensured consistent hydrological support for load estimation and validation across different spatial scales.

A more detailed description of the datasets used is provided in Sections 6.1 and 6.2 of the Tethys output O2.1.

The preparation of validation data was carried out through a structured, multi-step workflow implemented in R, combining water quality monitoring data with discharge information at the scale of the MoRE model analytical units (AUs). A detailed description of the individual processing steps, selection criteria, and calculation methods is provided in Section 6.3 of O2.1.

A key element of the validation approach was the careful spatial selection and assignment of monitoring stations to the analytical units used by the MoRE model. The monitoring stations were manually reviewed to ensure that selected points were representative of AU outlet conditions, with particular emphasis placed on the station location relative to catchment outlets. Where multiple stations were available within a single AU, preference was given to the most downstream station.

The spatial distribution of suitable monitoring stations proved to be a major limiting factor, especially in upstream catchments. In many cases, no station could be assigned to an AU despite the presence of multiple monitoring points within the catchment, due to suboptimal station placement relative to the outlet. Selection criteria were therefore applied more strictly in upstream areas, where representativeness is highly sensitive to station location, while criteria were gradually relaxed downstream along the Danube main stem, where cumulative effects dominate and outlet representativeness is less sensitive to small spatial offsets.

Within this framework, the inclusion of monitoring data from the Tethys project provided a clear added value for model validation. Tethys monitoring points significantly improved spatial coverage, particularly for pharmaceuticals and PFAS. In several regions, Tethys data enabled the assignment of validation points to AUs that previously lacked suitable monitoring stations.

This added value is illustrated by Figure 1, showing estimated average concentration values of diclofenac at validation points before and after the inclusion of Tethys monitoring data.

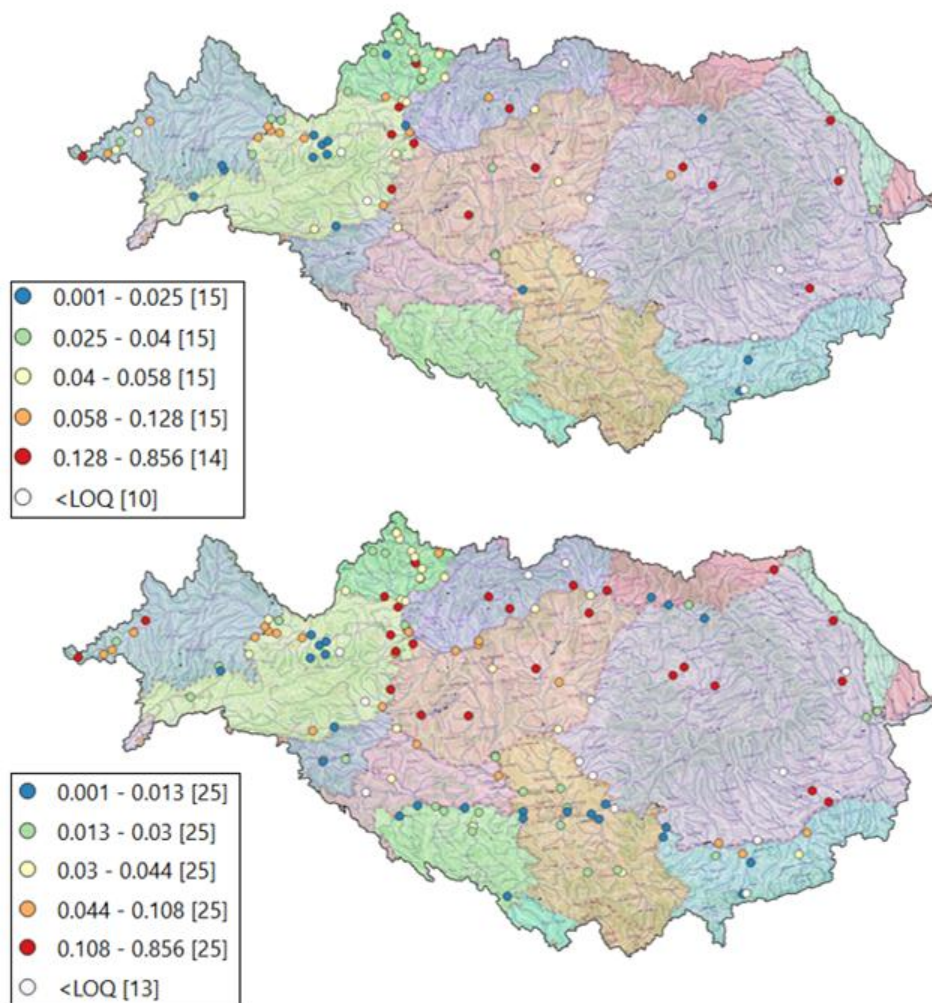


Figure 1: Average concentration – diclofenac [$\mu\text{g/L}$], before and after considering Tethys monitoring points

The process of selection of suitable monitoring stations was greatly supported by the transnational database and its PostGIS functionality, which allows direct integration of the database content (monitoring sites with their locations) into GIS software. The calculation of river loads benefits from having all necessary data for calculation of loads in a harmonized format available in one place.

2.7 Design of a pollution evaluation tool (PET)

The development of the pollution evaluation tool (PET) was kicked off at the data management workshop in Vienna, where team members from 10 partner institutions participated.

The aim of the workshop regarding the PET, was an open discussion among the project partners about the main aims and the key features, that the first version of the PET should contain. The following lists the agreement with descending importance:

- Visualization of concentration data including highlighting of outliers
- Improving the quality assessment workflow in the database in terms of speed and user experience
- Creation of merged and export ready data sets for e.g. modelling
- Completeness analysis of the imported data
- Visualization of data in general

During the design of the PET, certain demands were set mandatory to ensure a user-friendly and expandable tool for further use:

The entire application should be compactly stored in a project template structure³. Different topics are in individual files in their designated folders. This should preserve clear order, ensure a smooth workflow and enable easy distribution.

Codes should be clean, well-documented, robust, easy to expand, and structured into function chunks that perform one task each and can be used multiple times. In these functions the code should be wrapped in “tryCatch” functions which print unique error messages and should help significantly at error finding. This serves smooth understanding, expansion and error shooting of the code.

The code development platform Gitlab is applied to host the code and control the versioning, feature implementation and backup security in case of versioning issues or related compatibility errors. Her the tool is available for download:

<https://gitlab.tuwien.ac.at/e226-1-working-group-river-basin-management/tethys>

The information about every **Application**, **Package**, **Extension** or specified **Setting** (APES) used in this tool are documented in a README.txt file. As this tool will use several APES at the same time, a strict version control and compatibility tests are needed to ensure smooth development. To reduce the risk of errors due to incompatibility among APES, their quantity should be kept low if possible. Nevertheless, versioning will be mandatory.

With these rules set, the development started with the implementation of key functions of the tool regarding the smooth connection with the data base, the data transfer between the applications, the automated data aggregation, the plot and statistical functions as well as the automated report writing function.

These scripts stored in the “munge” directory build the fundament of the PET.

The config file was then filled with all relevant information regarding the APES of the scripts, mostly stored in in the global.txt file.

The scripts in the munge directory have the task to fill the files of the template accordingly as it is foreseen of the package and it is documented in the README.txt file.

In order to ensure a smooth start for new users the README.txt file explains the setup, the structure and best practice tips in ten steps leading to the three scripts long tutorial, that is stored in the file scr and explain all features the tool have.

The first script concerns itself with the import of data, the second with the plotting of data and the last with the creation of automated generated quality assessment reports of a data set.

After working through all scripts, the user is able to import and merge data sets fast, visualize a high variety of data of distinct origin and speed up the quality assessment process.

³ White J (2024)

3 Lessons learned and outlook

The pilot action showed already early in the process that the level of data literacy, meaning the capabilities in and knowledge about data collection, curation and evaluation is very heterogeneous among different institutions in the Danube basin. This is closely related to the availability of trained staff. While in some institutions personal responsible for chemical analyses in laboratories must handle data management as additional task, other institutions have dedicated experts for data management and modelling. This is on the one hand a challenge, as training must be organised considering very different levels of skills and knowledge. On the other hand, this means that by closely cooperating the partners can learn a lot from each other and help each other improving their data management, if this cooperation is organised in a sustainable way. By developing common tools and data formats, this cooperation can be strengthened and further exchange bilaterally among the countries and with the ICPDR as the transnational organisation can be simplified.

The ICPDR welcomes the development, testing and demonstration of the transnational database as a potential future operational tool at the basin-wide scale. The early involvement of and close coordination with the ICPDR experts ensured that the database is in line with the ICPDR needs, compatible with existing ICPDR data management tools and fully transparent and accessible for the ICPDR experts and stakeholders. The ICPDR will further explore how to adopt the database as an operational tool for future water management efforts on the basin-wide level.