Towards a water quality database for raw and validated data with emphasis on structured metadata

Queralt Plana, Janelcy Alferes, Kevin Fuks, Tobias Kraft, Thibaud Maruéjouls, Elena Torfs and Peter A. Vanrolleghem

ABSTRACT

On-line continuous monitoring of water bodies produces large quantities of high frequency data. Long-term quality control and applicability of these data require rigorous storage and documentation. To carry out these activities successfully, a database has to be built. Such a database should provide the simplicity to store and document all relevant data and should be easy to use for further data evaluation and interpretation. In this paper, a comprehensive database structure for water quality data is proposed. Its goal is to centralize the data, standardize their format, provide easy access, and, especially, document all relevant information (metadata) associated with the measurements in an efficient way. The emphasis on data documentation enables the provision of detailed information not only on the history of the measurements (e.g., where, how, when and by whom was the value measured) but also on the history of the equipment (e.g., sensor maintenance, calibration/validation history), personnel (e.g., experience), projects, sampling sites, etc. As such, the proposed database structure provides a robust and efficient tool for functional data storage and access, allowing future use of data collected at great expense.

Key words | big data, data management, data validation, filtering, SQL

Queralt Plana (corresponding author)
Janelcy Alferes
Kevin Fuks
Tobias Kraft
Thibaud Maruéjouls
Elena Torfs
Peter A. Vanrolleghem
modelEAU, Université Laval,
1065, avenue de la Médecine, Québec, QC, G1 V 0A6,
Canada
E-mail: *queralt.plana.1@ulaval.ca*

Queralt Plana
Elena Torfs
Peter A. Vanrolleghem
CentrEau, Quebec Water Research Centre, Université Laval,
1065, avenue de la Médecine, Québec, QC, G1 V 0A6,
Canada

Janelcy Alferes
s::can Messtechnik GmbH,
Brigittagasse 22-24, 1200 Vienna,
Austria

Tobias Kraft
AF Toscano AG,
Raetusstrasse 12, CH-7000 Chur,
Switzerland

Thibaud Maruéjouls
Le LyRE, Suez Eau France SAS,
Domaine du Haut-Carré 43, rue Pierre Noailles Bâtiment C4, 33400 Talence,
France

INTRODUCTION

Automated monitoring stations and state-of-the-art instrumentation are used to continuously monitor and control water bodies over the long term and increasingly also in real time. This on-line, continuous monitoring is used to collect data at high frequency thus generating large sets of data (Rieger & Vanrolleghem 2008). However, these large quantities of data are only beneficial if they are accessible, well-documented and reliable (Copp *et al.* 2010). Thus, the tasks of efficient storage and quality control are crucial to their interpretation and further application.

Generally, in many organizations, storage and quality check of the collected data are done individually by the users at their work space. However, each user organizes, structures and evaluates the data in a different manner (Camhy *et al.* 2012). As personnel are changing over time, this diversification hinders data interpretation,

doi: 10.2166/wqrj.2018.013

understanding and reproduction leading to inconsistencies in further studies.

Thus, to successfully manage these large amounts of heterogeneous data, a systematic and efficient storage system is needed (Rieger *et al.* 2004). In this respect, Camhy *et al.* (2012) and Horsburgh *et al.* (2008) identified several data management challenges: the collected raw data have a highly variable format; the database has to be flexible and adaptable because it is growing continuously: monitoring programs are modified, additional variables are measured and different sensors are used; the personnel involved in collecting and managing the data changes. It is thus critical that one is documenting the collected data with all relevant metadata (data about data).

Metadata are any additional information that provide more details about the data and its identification: the measured attributes, their names, units, the extent, the quality, the spatial and temporal aspects, the content, and how the value was obtained (Gray *et al.* 2005; ISO 2013). This information is essential for other potential users to understand and interpret the collected data.

The issues of metadata are illustrated with an example of a one-month measurement campaign conducted at a full-scale wastewater treatment plant. For this campaign, a number of automated sensors to measure water quality parameters (TSS, N-components, etc.) were installed. If only the measured values are stored, the data will only have very limited meaning. At the very least, metadata such as the variable names and their units should be stored as well. However, even with the addition of these metadata, the relevance and application of the data set will most likely be limited to persons that were directly involved in the campaign. Subsequently, the data will either be shelved and lost or applied unsuccessfully in a further study because too much information on the data is missing. If we want the efforts of such a measurement campaign to transcend this limited life-expectancy, much more detailed metadata should be stored: the exact location where the sensors were placed, the type of sensors (and their measurement principles), their maintenance, calibration and validation history, the weather conditions during the campaign, etc. Providing a systematic structure to store all these metadata is an important challenge for effective data management.

Some commercial databases to store water quality and hydrological data in a structured way are offered on the market. Nevertheless, accessing the raw data or making a modification of the metadata is sometimes limited or not possible, and can only be done through a predefined graphical user interface (GUI) (Camhy *et al.* 2012). Moreover, data have to be continuously transformed to the proprietary format of the software. In addition, any modification relies on the vendor support, thus placing important restraints on customized use.

Also, some organizations have proposed standards to exchange environmental data including data description, analysis and reporting, e.g., the Environmental Data Standards Council (EDSC) presented a manual on Environmental Sampling, Analysis and Results Data Standards (EDSC 2006), the National Water Quality Monitoring Council (NWQMC) developed a similar standard but specific for water quality (NWQMC 2006), and the Open Geospatial Consortium (OGC) presented the 'Observations and Measurements' best practices document (OGC 2006). Despite that, these standards are focused on the elements to transfer and exchange the data rather than how to structure the data in a relational database.

In recent years, some hydrological and water quality databases have been developed, e.g., the Observations Data Model (ODM) database from the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) (Horsburgh *et al.* 2008), or the STOrage and RETrieval (STORET) database developed by the US Environmental Protection Agency (EPA) (EPA 2016). However, storage and access to metadata is still a challenge. Most of the published databases focus on measurement and location details, providing priority to data collection activities and data set characteristics rather than information about monitoring programs. Moreover, some limitations are also observed on the control of the data quality (Horsburgh *et al.* 2015).

Using their experience with high frequency data collection, the model*EAU* research group at Université Laval in Québec City (Canada), developed a database structure to be applied to water quality data from rivers, sewer systems and water resource recovery facilities (WRRFs). The main objectives of this database are to centralize data storage from on-line measurements, laboratory analysis and data post-treatments, and deal with the challenges presented above, especially regarding the storage of metadata. This

paper presents the structure of the developed database and its application.

DATABASE DESIGN

The database structure that was designed, named dat*EAU*base (water database, 'eau' is water in French), offers robustness, data format uniformity, flexibility if modifications are needed, efficient storage of relevant metadata, and the possibility to comprehensively document a monitoring program.

The dat*EAU*base has been designed to store all relevant data, i.e., the raw, filtered and validated data, laboratory measurements and corresponding metadata (see Figure 1). The storage of the raw, filtered and laboratory data in the same database has been considered essential since all of them are related, and crucial to validate the data series and assure their quality.

dat*EAU*base STRUCTURE

The metadata considered are presented in Figure 2 and include detailed information about the sites, the sampling points, the watershed, the parameters, the equipment used, the measurement procedure followed, the project in which the data have been collected, for which purpose the value has been measured, the person responsible for the value and the weather conditions when the value was taken.

The design presented in Figure 2 is materialized by 23 different, interrelated tables in MySQL. The overall structure of the dat*EAU*base is presented in Figure 3. Compared to other software, e.g., MS Access, MySQL not only offers a large capacity but, more importantly, also the possibility to work with m-to-n relationships (MS Access for instance, only allows 1-to-n relations). The m-to-n relationship means that each row in one table can be related to multiple rows in another table and vice versa. For example, many people can be involved in one project, and one person can also be involved in several projects. The links between the tables are made through the specific keys (called IDs in Figure 3) associated with each row of a table. The storage requirements for each data type included in the data*EAU*base are described in Table 1.

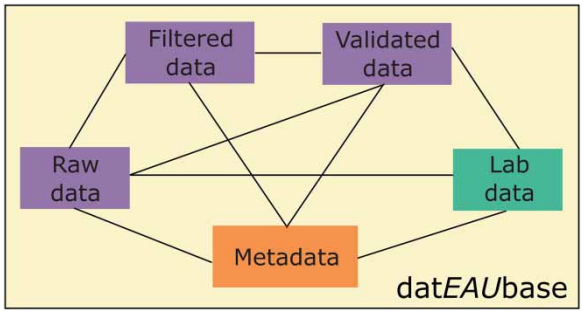

Figure 1 | Modular design of the datEAUbase.

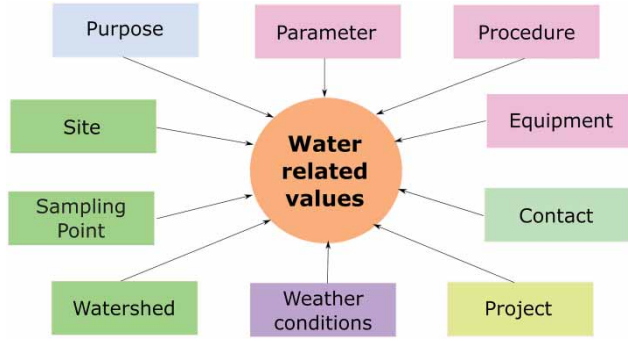

Figure 2 | dat*EAU*base structure.

Primary tables

The general structure is based on primary and lookup tables. The primary tables (*Metadata*, *Value* and *Comments* tables presented in Figure 3) are the main tables of the database. Each measured value, its corresponding time stamp and its replicate identification (under 'Number_of_experiment') are stored in the *Value* table. Through its Metadata_ID each value is linked to a specific set of metadata in the *Metadata* table. Moreover, any comment can be added next to a value if needed.

The *Metadata* table contains a list of all existing metadata combinations. This list only consists of IDs that represent links to more detailed information in the lookup tables. Hence, the *Metadata* table is directly or indirectly linked to all other lookup tables (see Table 2).

To illustrate the database's structure, an example follows. In the primary tables, the information stored can be: on June 15, 2015 at 10:40:00 GMT, a value of 6.5 was measured. This value is linked to Metadata_ID 22. Moreover, a comment can be added that the calibration activity was unsuccessful. Through the internal links with the

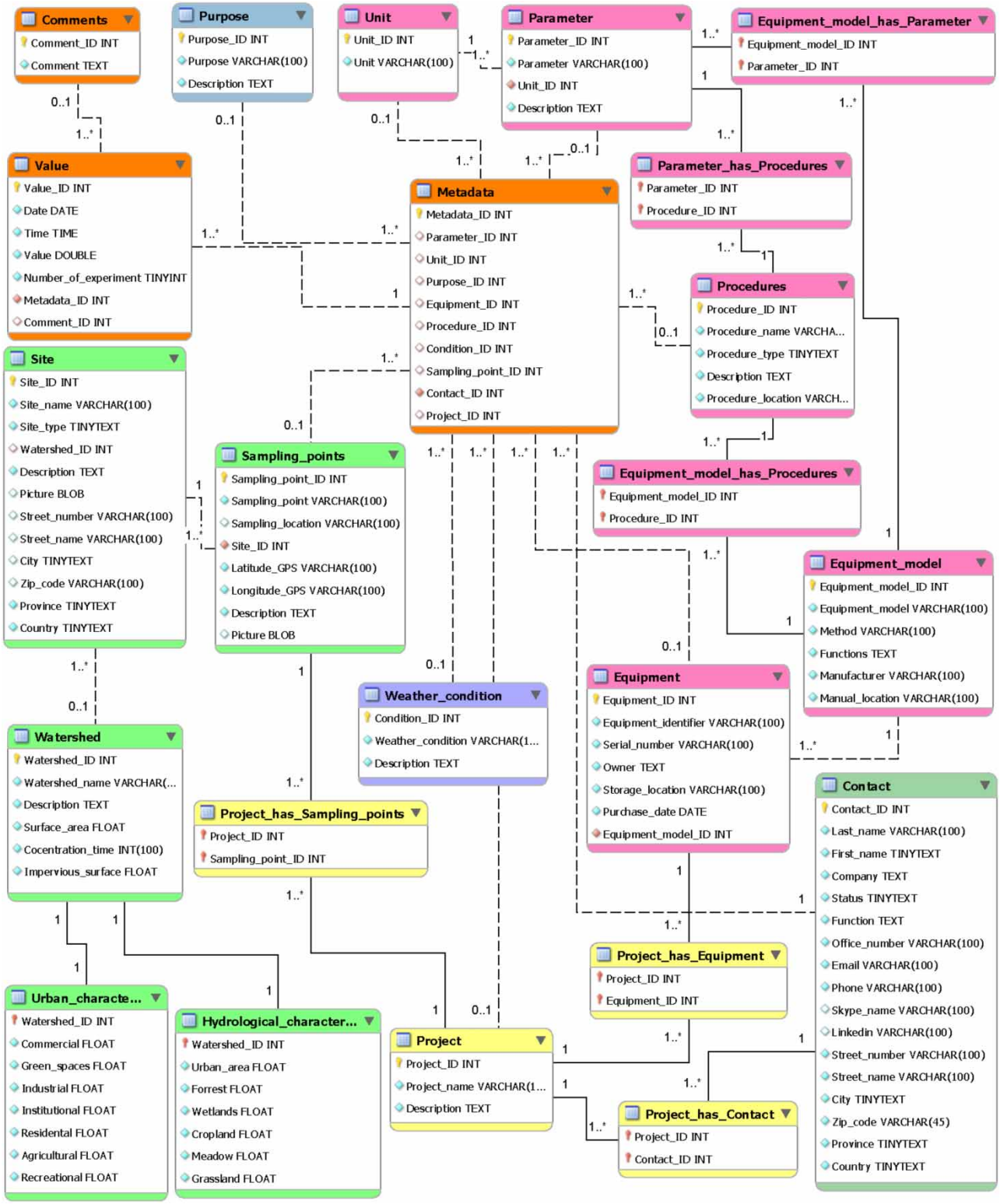

Figure 3 | dat*EAU*base model with the links between the tables. The primary keys of each table are designated with a key, all diamonds represent foreign keys.

Table 1 | Data storage requirements for data included into the datEAUbase

Data type	Storage required	Description
TINYINT	1 byte	A very small integer
INT	4 bytes	A normal-size integer
FLOAT	4 bytes	A small (single-precision) floating-point number
DOUBLE	8 bytes	A normal-size (double-precision) floating-point number
DATE	3 bytes	A date
TIME	3 bytes	A time
TINYTEXT	$L + 1$ byte, where $L < 2^8$	A TEXT column with a maximum length of 255 $(28 - 1)$ characters
TEXT	$L + 2$ byte, where $L < 2^{16}$	A TEXT column with a maximum length of 65,535 $(216 - 1)$ characters
VARCHAR (100)	$L + 1$ byte	A variable-length string
BLOB	$L + 2$ byte, where $L < 2^{16}$	A text as the corresponding binary string data type

Table 2 | Information included in the metadata table

Table columns	Characteristic	Description
Metadata_ID	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Parameter_ID	Foreign Key	Measured parameter. Link to the *Parameter* table
Unit_ID	Foreign Key	Unit of the parameter. Link to the *Unit* table
Purpose_ID	Foreign Key	Purpose of the data collection. For example: Measurement, laboratory analysis, calibration or cleaning. Link to the *Purpose* table
Equipment_ID	Foreign Key	Equipment which was used. Link to the *Equipment* table
Procedure_ID	Foreign Key	Procedure corresponding to the purpose and/or the equipment. Link to the *Procedure* table
Condition_ID	Foreign Key	Weather condition during the measurement. Link to the *Weather_condition* table
Sampling point_ID	Foreign Key	Sampling point where the data were collected. Link to the *Sampling_point* table
Contact_ID	Foreign Key, not null	Person who is responsible for the measurement. Link to the *Contact* table
Project_ID	Foreign Key	Name of the project for which the data was collected. Link to the *Project* table

different lookup tables, Metadata_ID 22 can be translated to a measure of pH, which has no units, with the sensor pH_003 under dry weather conditions, with the purpose of calibrating the sensor according to the ISO-15839 methodology, at the inlet of the Grandes-Piles facultative aerated lagoon (F/AL) by Plana for the mon*EAU* project. More information on the measurement principle of the pH_003 sensor, the location of the Grand-Piles facultative aerated lagoon or the mon*EAU* project can then be found in the corresponding lookup tables.

The use of the lookup tables together with the links between the tables, especially the n-to-m links, allows for very efficient storage of huge amounts of information by avoiding redundancy. For example, information on a certain equipment model has to be stored only once, then every equipment of this model is directly linked to this piece of information. Also, the equipment model is directly linked to one or more parameters, but the equipment itself is not as this would create a triangular relationship. Finally, each measured value is linked to a certain combination of existing metadata through the *Metadata* table. Since this table only consists of IDs (i.e., integers), the storage volume is highly reduced. In this way, once the metadata information is loaded into the dat*EAU*base, storage needs are defined only by the storage of the Value Table, i.e., a timestamp, a double and a Metadata_ID. For the current set-up of the dat*EAU*base in model*EAU*'s laboratory this comes down to approximately 0.1 kB per datapoint. With an average of 150,000 online values stored per day, current storage requirements are approximately 20 MB/day. On a monthly basis this represents 0.5 Gb.

Ultimately, by its specific structure the dat*EAU*base not only permits to rigorously document all measured values but it also allows to build memory of the measuring campaigns in a reliable way. For instance, the structure allows to track the history of a piece of equipment, e.g., in which projects has one sensor been used or which is its calibration/

validation history; the history of the personnel is also tracked, e.g., who has been involved in a certain project or who has used certain equipment which can be useful information if some experienced person is needed.

Lookup tables

The lookup tables have been divided into six different blocks, shown in Figure 3: all information about the instrumentation is stored in *Equipment*, *Equipment_model*, *Procedures*, *Parameter* and *Unit* tables; the information about the sampling point is stored in *Site*, *Sampling_points*, *Watershed*, *Urban_characteristics* and *Hydrological_characteristics* tables; the project information is stored in the *Project* table; the information of the people involved is stored in the *Contact* table; the purpose of the measurement is stored in the *Purpose* table; and the weather information is stored in the *Weather_condition* table.

Instrumentation information

The set of tables related to instrumentation provides detailed information about the equipment and measurement procedures, as well as which parameters can be measured with the equipment and the units used.

For example, taking the parameter measured to be pH, it first of all has no units. It is measured with the sensor pH_003 corresponding to the Hach's model DPD1P1 with the serial number 2659777. The measurement principle of this sensor is a differential of the electrical potential. For further information about the sensor, its manual can be found at location PLT-2659. Currently, the sensor is installed at the Grandes-Piles F/AL for on-line measurement. For a proper maintenance, standard operating procedure SOP_49_pH should be followed which is also stored in room PLT-2659.

Sampling location information

The *Sampling location* tables contain the information about the site and the identification of the specific sampling points. Also, some more information about urban and hydrological characteristics is included.

For example, measurements are collected at the inlet of the Grandes-Piles F/AL. This F/AL's address is 267–303 5e Av., Grandes-Piles, G0X 1H0, QC, Canada and the specific coordinates at the inlet are 46°41′04″N 72°42′59″W. The watershed of this location is the Saint-Maurice river with a surface area of 43,300 km^2. The concentration time of this watershed is 2 days and the impervious surface is about 4%. Its urban characteristics are 54.25% of green space, 2.25% of industrial area, 13.5% of residential area, 22% of agricultural area and 8% of recreational area. Its hydrological characteristics are 17% of urban area, 39% of wetlands, 12% of croplands, 8% of meadow and 3% of grassland.

Project information

In the *Project* table, information about the project is detailed. This table is linked to other parts of the database by a number of tables containing n-to-m links. These linking tables contain information about who is working in a project, where a project takes place and which equipment is used, and vice versa, in how many projects someone is working, for how many projects a location is used, and in how many projects a piece of equipment is used.

For example, the mon*EAU* project deals with the usefulness of automatic monitoring stations (AMS) to study the water quality. The measurements are located at the inlet of Grandes-Piles F/AL. The following equipment is used: conductivity_001, pH_003 and ammolyser_001. The personnel involved are Alferes, Plana and Vanrolleghem.

Contact information

In the *Contact* table, detailed information about the people involved in the different projects is stored. This information includes the first name, the last name, their affiliation together with the address of the corresponding office and the person's function. Also, the e-mail, the phone number, the skype name or the LinkedIn information are stored.

Purpose of the measurement information

The *Purpose* table stores information about the aim of the value included in the database, i.e., on-line measurement, laboratory analysis, calibration, validation or cleaning. This

is accompanied with a detailed description of the different purposes.

For example, the purpose of the measurement is sensor validation. This is a routine sensor validation activity for verification of proper operation.

Weather information

Despite the fact that weather data such as daily rainfall or hourly temperatures can be stored into the database, this table allows to link directly to the measured value of any parameter information on such characteristics as dry weather, wet weather or snow melt. For example, wet weather conditions are considered to have rainfall of more than 3 mm/d.

datEAUbase APPLICATION

The structure and design of the datEAUbase creates a comprehensive environment to store and document data alongside their relevant metadata in a robust and highly efficient way. Moreover, it ensures that each value stored in the datEAUbase is unique, being linked to a specific time stamp and a complete set of metadata.

Although these features represent the core functionality of the datEAUbase, in reality such a large-scale database will only be useful if the information contained within is easily accessible for all users. Hence, tools should be in place to facilitate interaction with the database. External interaction with the datEAUbase currently consists of two different parts (see Figure 4): automatic read-in of online data from data loggers and a user-friendly web interface (programmed in PHP) which allows further manual entry as well as a comprehensive search, viewing and export of the stored data.

Data transmission from on-line monitoring stations into the datEAUbase is ensured through a secure VPN connection. The actual upload procedure depends on the specific properties of the monitoring station and more specifically on the raw data format of the sensor data. For the current application, MSSQL, Python and Visual Basic programs are available to handle the most common data formats, i.e.: SQL tables, data text files and binary data files. Through these different scripts continuous upload of on-line measurement data into the datEAUbase is established.

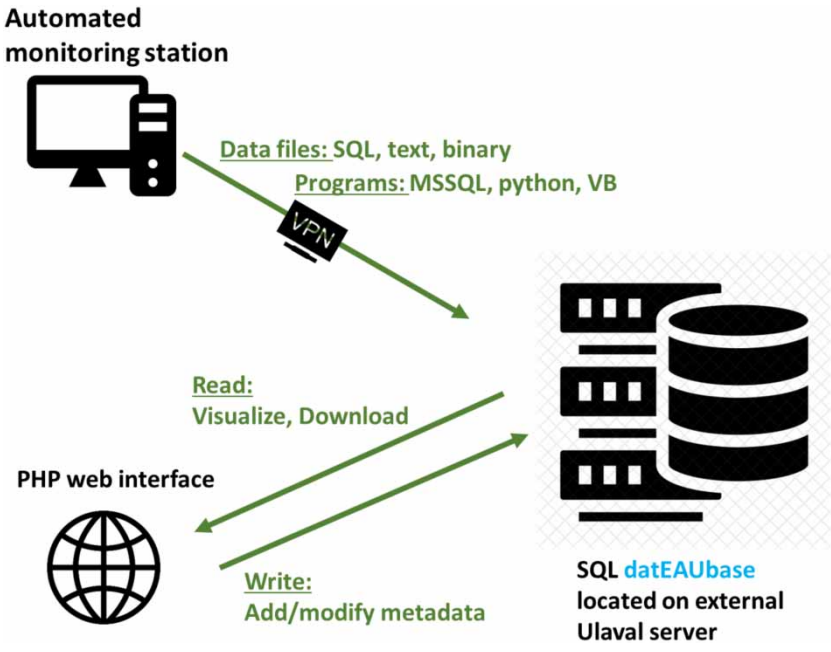

Figure 4 | Data flow design of the datEAUbase.

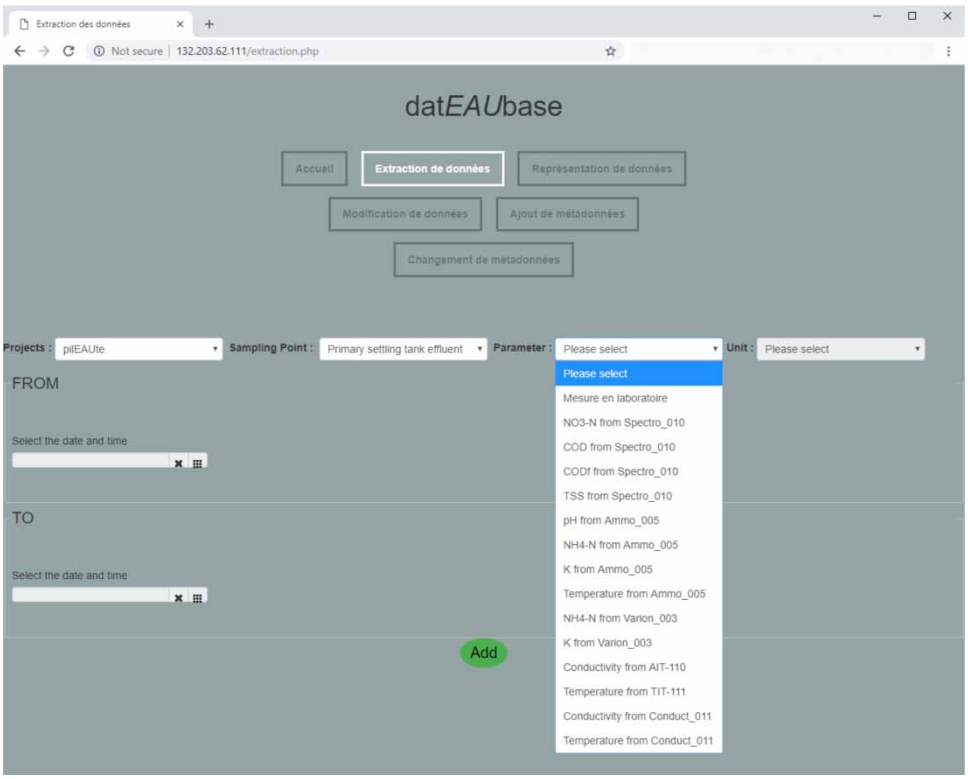

Figure 5 | Screenshot of the dat*EAU*base interface application.

The following important steps in the maintenance and application of the dat*EAU*base are facilitated through the user interface (Figure 5):

- Before measurements can be stored in the dat*EAU*base, its metadata need to be present in the lookup tables. The interface allows easy addition or modification of metadata (for example, adding a new sensor in an existing project).
- Different metadata_IDs have to be created in the metadata table for all existing metadata combinations. Such changes to the metadata table do not occur continuously but are associated with well-defined events (e.g., when a new sensor is bought, a sensor is relocated, a new project is started). The interface allows an easy check as to whether a certain combination of metadata is already present in the database or if it should be created. Once the metadata_ID for an online sensor is created, online data from this sensor can easily be stored in the database through coupling with its metadata_ID.
- Non-automated data (such as laboratory results) can be entered in the dat*EAU*base through the user interface. This also consists of a simple coupling of the measured values to their corresponding metadata_ID.
- One of the main features of the interface is its application to search the database and extract a specific data set of interest or information on sensor or project history.
- During the search process, an internal quality check is also performed. Data will only be available for extraction if all internal links are present. All metadata combinations that are present in the metadata table should also be linked internally in the lookup tables.

CONCLUSIONS

Technological advances in water quality measurement lead to the creation of large quantities of high frequency data. Without efficient storage and rigorous documentation, the

life expectancy of these data is often limited to the specific project for which they were collected. Such common practices represent a significant loss of information as well as expense (that often goes into a measurement campaign). To maintain understanding of the collected data, track their history and secure their usefulness in further studies, documentation by metadata is crucial. This includes detailed information about the sites, the sampling points, the watershed, the parameters, the equipment used, the measurement procedure followed, the project in which the data have been collected, for which purpose the value has been measured, the person responsible for the value and the weather conditions when the value was taken.

This paper presents a comprehensive database structure (the dat*EAU*base) that offers a data storage system with an emphasis on metadata. It provides a robust, large storage capacity with flexibility for future modifications and possible improvements.

Its specific structure, consisting of a combination of three primary tables interlinked with 20 lookup tables, allows for very efficient storage of huge amounts of information while avoiding redundancy. Moreover, this rigorous documentation of all measured values with their metadata allows to build memory on sensor history, project history and so on, in a reliable way.

Since this tool is meant for large data users to store and exchange water quality data, easy access and maintenance is ensured through a user-friendly interface.

ACKNOWLEDGEMENTS

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant, the Urban Science Joint Research Unit (Unité Mixte de Recherche en sciences urbaines, UMRsu), the Quebec consortium for industrial bioprocess research and innovation (CRIBIQ), and MITACS. Peter A. Vanrolleghem holds the Canada Research Chair in Water Quality Modelling.

REFERENCES

Camhy, D., Gamerith, V., Steffelbauer, D., Muschalla, D. & Gruber, G. 2012 Scientific data management with open source tools – An urban drainage example. In: *Proceedings IWA/IAHR 9th International Conference on Urban Drainage Modelling*, September 4–6, 2012, Belgrade, Serbia.

Copp, J., Belia, E., Hubner, C., Thron, M., Vanrolleghem, P. A. & Rieger, L. 2010 Towards the automation of water quality monitoring networks. In: *Proceedings Automation Science and Engineering (CASE), IEEE*, August 21–24, 2010, Toronto, Ontario, Canada, pp. 491–496.

EDSC 2006 *Environmental Sampling, Analysis and Results Data Standards*. Environmental Data Standards Council (EDSC), US Environmental Protection Agency, Washington, DC, USA.

EPA 2016 *STOrage and RETrieval Data Warehouse*. US Environmental Protection Agency. https://www.epa.gov/waterdata (accessed 13 December 2016).

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J. & Heber, G. 2005 Scientific data management in the coming decade. *ACM SIGMOD Record* **34**, 34–41.

Horsburgh, J. S., Tarboton, D. G., Maidment, D. R. & Zaslavsky, I. 2008 A relational model for environmental and water resources data. *Water Resources Research* **44**, 1–12.

Horsburgh, J. S., Reeder, S. L., Jones, A. S. & Meline, J. 2015 Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environmental Modelling & Software* **70**, 32–44.

ISO 2003 *ISO 19115:2013 Geographic Information – Metadata*. International Organizations for Standardization, Geneva, Switzerland.

NWQMC 2006 *Water Quality Data Elements: A User Guide*. National Water Quality Monitoring Council (NWQMC), Washington, DC, USA.

OGC 2006 *Observations and Measurements*. Open Geospatial Consortium (OGC), Wayland, MA, USA.

Rieger, L. & Vanrolleghem, P. A. 2008 mon*EAU*: a platform for water quality monitoring networks. *Water Science and Technology* **57** (7), 1079–1086.

Rieger, L., Thomann, M., Joss, A., Gujer, W. & Siegrist, H. 2004 Computer-aided monitoring and operation of continuous measuring devices. *Water Science and Technology* **50** (11), 31–39.

First received 2 March 2018; accepted in revised form 15 October 2018. Available online 14 November 2018