dat*EAU*base: Water quality database for raw and validated data with emphasis on structured metadata

Queralt Plana[1]*, Janelcy Alferes[1,2], Kevin Fuks[1], Tobias Kraft[1], Thibaud Maruéjouls[1,3], Elena Torfs[1] and Peter A. Vanrolleghem[1]

[1]model*EAU*, Université Laval, 1065, Avenue de la Médecine, Québec (QC), G1V 0A6, Canada

[2]SUEZ environnement SAS, chemin de ronde 87, 78290 Croissy sur Seine, France

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

*Corresponding author: *queralt.plana.1@ulaval.ca*

Abstract: On-line continuous monitoring of water bodies produces large quantities of high frequency data. Long-term quality control and applicability of these data requires rigorous storage and documentation. To carry out these activities successfully, a database has to be built. Such 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 presented. Its goal is to centralise 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 allows to provide 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... As such, the presented database provides a robust and efficient tool for functional data storage and access, allowing future use of the data collected at great expense.

Keywords: Big data, SQL, filtering, data validation, data management

INTRODUCTION

Automated monitoring stations and state-of-the-art instrumentation are used to continuously monitor and control water bodies over 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 and 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 organises, structures and evaluates the data in a different manner (Camhy et al., 2012). Adding that personnel is changing over time, this diversification hinders data interpretation, 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 to collect and manage the data changes. It is thus critical that one is documenting the collected data with all relevant metadata (data about data).

Metadata is any additional information that provides 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 (ISO, 2013; Gray et al., 2005). This information is essential for other potential users to understand and interpret the collected data.

The issues of metadata illustrated are 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...) 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 dataset 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... 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 organisations 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 U.S. 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, 2015).

Considering their experience with high frequency data collection, the model*EAU* research group at Université Laval in Québec City (Canada), developed an internal database applied to water quality data from rivers, sewer systems and water resource recovery facilities (WRRFs). The main objectives of this database are to centralise data storage from on-line measurements, lab analysis and 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 the monitoring program.

The dat*EAU*base has been designed to store all relevant data, i.e. the raw, filtered and validated data, lab measurements and corresponding metadata (See figure 1). The storage of the raw, filtered and lab 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.

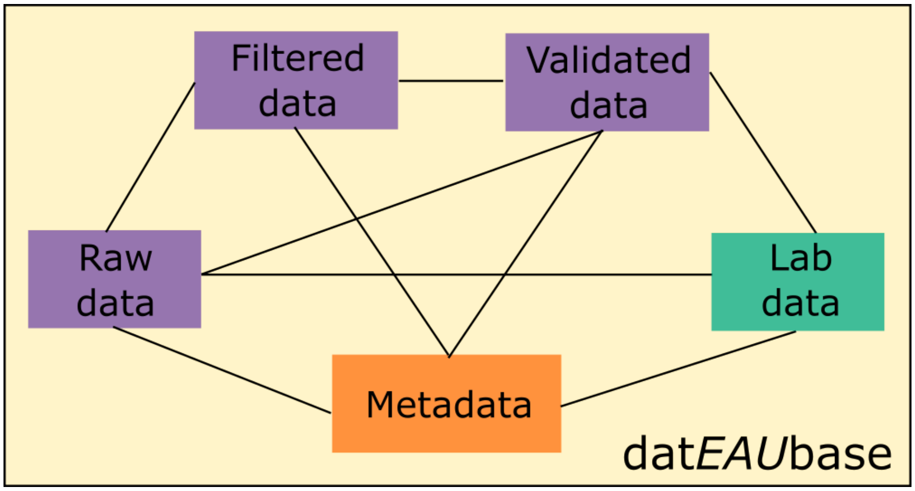

Figure 1. Modular design of the dat*EAU*bsae.

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 of 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. Following the same color scheme, the overall structure of the dat*EAU*base is presented in figure 3. Compared to other softwares, e.g. MS Access, MySQL offers a large capacity but, more importantly,

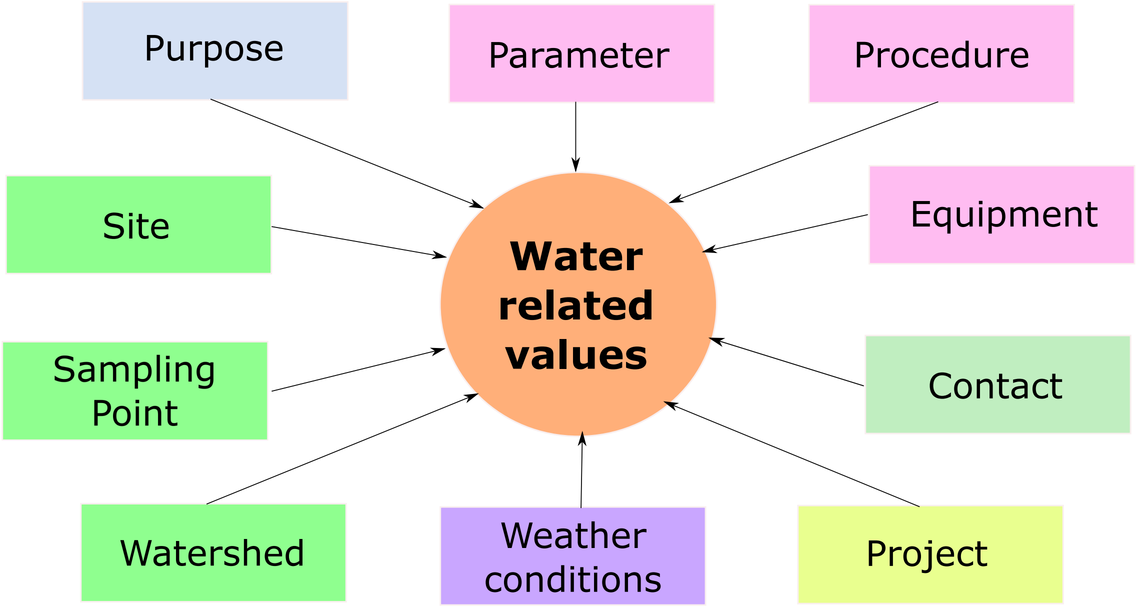

Figure 2. dat*EAU*base structure.

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.

Primary tables

The general structure is based on primary and lookup tables. The primary tables (presented in orange in figure 3) are the main tables of the d atabase. Each measured value and its corresponding time stamp is 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 1).

To illustrate the database's working, 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 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 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 aereated lagoon or the mon*EAU* project can then be found in the corresponding lookup tables.

Table 1. Information included into 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, lab analysis, calibration or cleaning. Link to a 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_conditio*n table
Sampling_point_ID	Foreign Key	Sampling point where the data was collected. Link to the *Sampling_point* table
Contact_ID	Foreign Key, not null	Person who is responsible of 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

The use of the lookup tables together with the links between the tables, especially the n to m links, allow 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.

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 in six different blocks. Each block has been identified by different colors in figure 3: all information about the instrumentation is stored in the pink tables; the information about the sampling point is stored in the green tables; the project information is stored in the yellow tables; the information of the people involved is stored in the dark green table; the

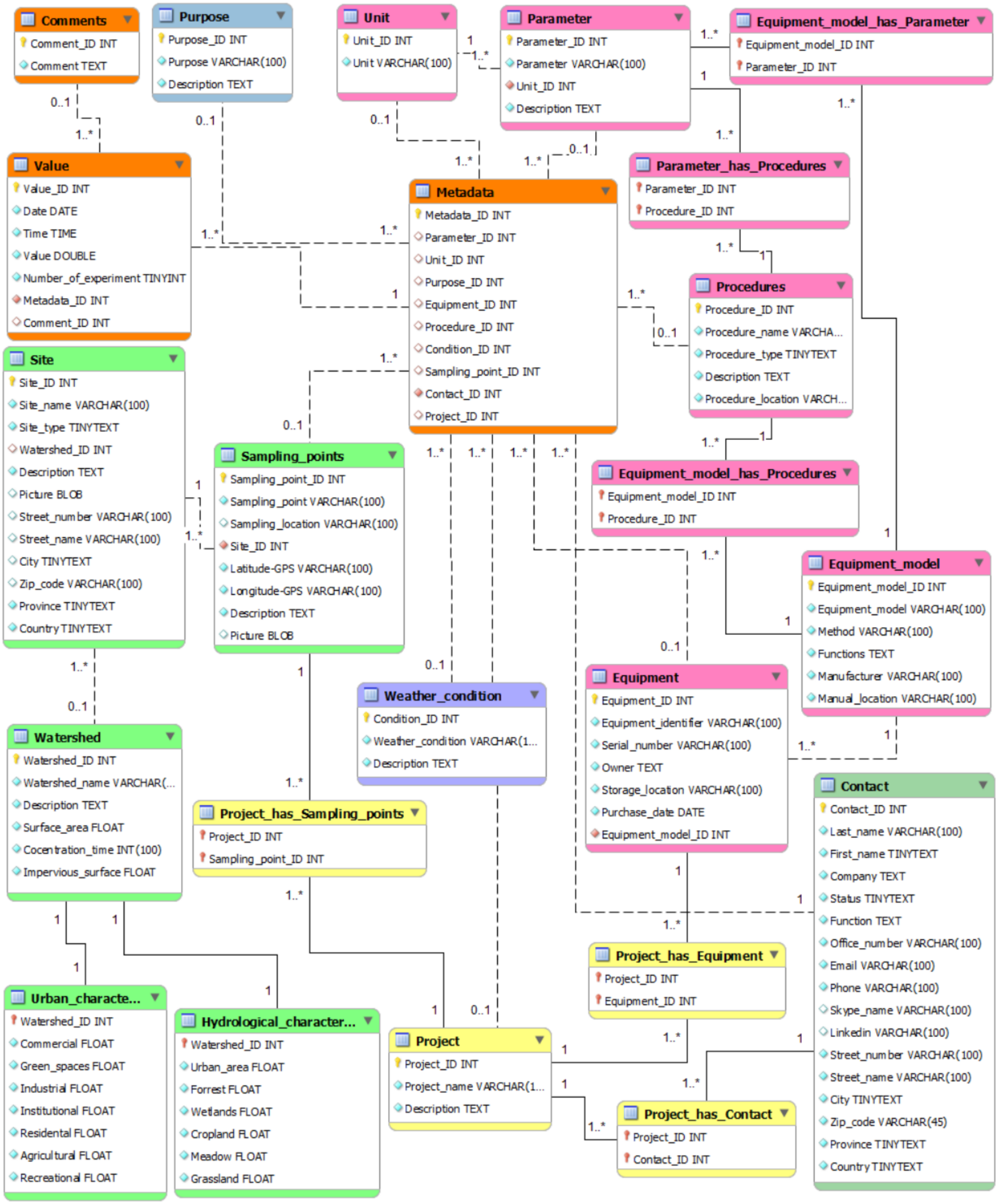

Figure 3. dat*EAU*base model with the links between the tables. The primary keys of each table are designated with a key, all diamonds represent foreign keys.

purpose of the measurement is stored in the blue table; and the weather information is stored in the purple 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.

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: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. And, the personnel involved is 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.

For example, Queralt Plana, PhD student at Université Laval, 1065, Avenue de la Médecine, local 2954, Québec (QC), G1V 0A6, Canada. Her work phone number is +1(418)656-2131, ext. 8730 and her e-mail address queralt.plana.1@ulaval.ca.

Purpose of the measurement information. The *Purpose* table stores information about the aim of the value included into the database, i.e. on-line measurement, lab 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.

dat*EAU*base APPLICATION

The specific structure and design of the dat*EAU*base 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 dat*EAU*base is unique, being linked to a specific time stamp and a complete set of metadata.

Although these features represent the core functionality of the dat*EAU*base, 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 dat*EAU*base currently consists of two different parts (See Figure 4): automatic read-in of online data from data loggers and a user-friendly interface which allows further manual entry as well as a comprehensive search, viewing and export of the stored data.

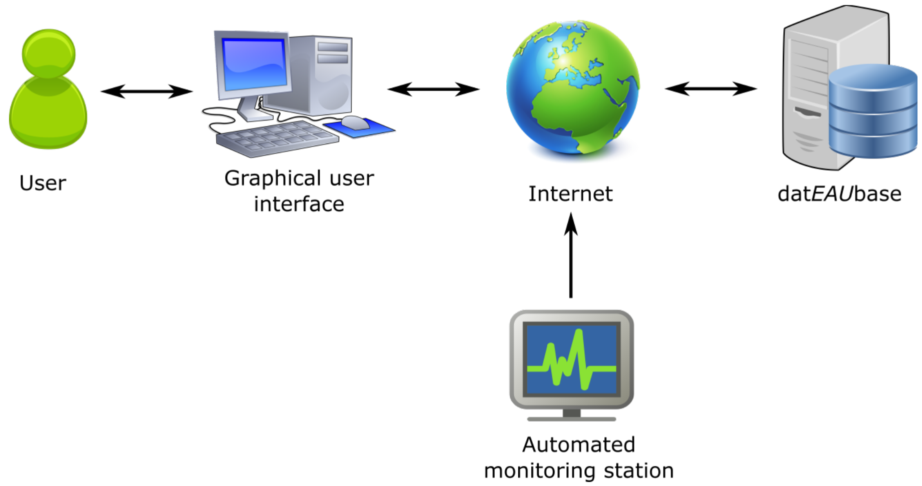

Figure 4. Data flow design of the dat*EAU*base.

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

- 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 to 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 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 lab results) can be entered in the dat*EAU*base through the user interface. Also here, this 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 dataset 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 frequent 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 maintain 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 of the value and the weather conditions when the value was taken.

This paper presented 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 3 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

APHA (1995). *Standard Methods for the Examination of Water and Wastewater*. 19th edn, American Public Health Association/American Water Works Association/Water Environment Federation, Washington DC, USA.

Camhy, D., Gamerith, V., Steffelbauer, D., Muschalla, D. and 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, Belgrade, Serbia, September 4-6 2012*.

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

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

EPA (2016). STOrage and RETrieval Data Warehouse. https://www.epa.gov/waterdata. U.S. Environmental Protection Agency. Accessed: 2016-12-13.

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

Horsburgh, J. S., Reeder, S. L., Jones, A. S. and 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.

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

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

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