Data communication and storage

Flora Branger, Simon Tait, Véronique Chaffard, Elodie Brelot, Vivien Lecomte, Isabelle Mallet, Peter Skipworth

Chapter 5

Data communication and storage

Flora Branger[1], Simon Tait[2], Véronique Chaffard[3], Elodie Brelot[4], Vivien Lecomte[4], Isabelle Mallet[5] and Peter Skipworth[6]

[1]INRAE, UR Riverly, Villeurbanne, France
[2]University of Sheffield, Civil and Structural Engineering, Sheffield, UK
[3]IRD – Institut de Recherche pour le Développement, IGE, Grenoble, France
[4]GRAIE, Villeurbanne, France
[5]Métropole de Lyon, Lyon, France
[6]Environmental Monitoring Solutions Ltd., Sheffield, UK

ABSTRACT

Monitoring programs in urban drainage systems generate, potentially, a huge amount of data from sources distributed in the urban environment, working at relatively high sampling rates for extended periods of time. Collecting data using adaptable and reliable communication systems is the first challenge. Then structuring the collected data is a first requisite for effectively managing the quality and accessibility of the data. In adjacent fields of research, the topic of managing huge collections of data has resulted in several (open) standards and protocols for database structure, transfer and storage to ensure unambiguous definitions on which parties can build their workflows/software. This chapter describes relevant approaches for urban drainage and stormwater management systems, and appropriate standards along with examples from case studies.

Keywords: Data accessibility, database, data communication, data standards.

doi: 10.2166/9781789060119_0135

5.1 INTRODUCTION

The cost of reliable data collection is becoming a significant burden for many water utilities and a major constraint for researchers in the field of urban drainage and stormwater management (UDSM). For water utilities there is often a need to demonstrate to a government regulator that a particular level of performance has been achieved. This need is likely to increase as monitoring is a key aspect of the Water Framework Directive (European Union [EU], 2000) and requires demonstration of compliance with earlier legislation such as the Urban WasteWater Treatment Directive (EU, 1991, 1998). An example of this need for enhanced monitoring is the instruction sent to UK water companies in September 2013 requiring that the majority of combined sewer overflows (CSOs) must be monitored for their frequency and duration of operation. Whilst the measurement itself may have its own complexity (see Chapters 3 and 6), the need to transfer the data securely, the number of sites involved and the need to demonstrate the quality of the collected data to the environmental regulator made this a technically challenging task. Currently the regulator, the UK's Environment Agency, has taken a risk-based approach to implement this monitoring program. It is expected that by 2020 over 30,000 CSOs in the UK will have event-based monitoring in which the data are telemetered back to central sites or that data are collected and stored securely before being collected manually. This example is a good illustration of the costs of data storage and communication for system monitoring for environmental purposes.

With various quantities measured at many locations, with, in general, consistent frequency data collection needed, for example every minute or so (see Section 6.2.5) to be able to capture the dynamic behaviour of an urban drainage system, and over long periods of time of the order of many months, data sets can quickly become overwhelmingly large. The challenges of data communication and storage are therefore not losing any information and ensuring that the data can be efficiently accessed and used, whatever that use might be (data analysis, calibration of numerical models, publication, etc.). This requires a carefully thought workflow from the sensor to the database, and appropriate strategies for data structuration and storage. Another challenge is that there is also increasing pressure from public authorities to make data available for the public. For example, the new European Open Data Directive (EU, 2019) requires from the Member States that all data produced by public sector activities (including research, ministries, state agencies and municipalities) should be available with open access or at marginal costs, and should be reused as much as possible, including for economic purposes. This goes

Key messages on data communication and storage and workflow in urban drainage systems

- KM 5.1: *Data Communication* – Two key issues need to be considered: reliability of data transfer and data latency. Acceptable threshold of reliability and latency are determined by the intended use of the data, for example if the data is to be used for monitoring or warning, each use requires different minimum levels of reliability and latency. Cost and power usage are secondary concerns.
- KM 5.2: *Data storage* – The use of data base management systems requires an important initial investment but guarantees data security, fast and efficient access and opportunities for data sharing and interoperability in the long-term. However tempting, spreadsheets are no DBMS (Data Base Management System) and are not appropriate for data storage.
- KM 5.3: *Workflow* – The workflow process for supplying data to the database and updating it regularly is as important as the design of the database structure itself.

along with the concept of FAIR data that is now becoming a major objective of data management in research and many public administrations (Wilkinson *et al.*, 2016). FAIR means Findable, Accessible, Interoperable and Reusable. These are the conditions that must be met in order to make data available in a long-term perspective and taking advantage of the current advanced technologies (cloud storage, webservices, big data). The development of a communication and storage strategy should therefore also take that into account.

This chapter presents the current state of the art and guidelines for building a strategy and choosing the appropriate tools for data storage and communication, from *in situ* sensors to interoperable databases. It also provides practical case studies with recommendations from practitioners.

5.2 FROM *IN SITU* SENSORS TO DATA FILES – DATA TRANSFER METHODS

Collecting spatially distributed real time data in urban drainage and stormwater management systems is challenging. The traditional way in which researchers and water authorities have transferred data is using SCADA (supervisory control and data acquisition) systems. This approach uses technology that was developed in the 1980s. This traditional approach requires significant investment and delivers systems with inflexibility and significant ongoing costs and is often unsuitable for research studies. More recently many commercial sensor units have coupled to GSM (global system for mobile communication) communication modules, but these often have issues with data reliability, mainly caused by the intermittence of the GSM network and poor power reliability. Wireless communication between sensors and local hubs via radio telemetry is becoming more popular due to its lower costs and reduced latency and reliability issues compared with cellular communication. Radio telemetry is a low power technology. Transmitters and receivers can thus be developed with long battery lives of up to 5 years. Some radio devices can be set to transmit at pre-set intervals, increasing the transmit frequency if certain urban drainage conditions are met, such as high-water levels.

The industrial, scientific and medical (ISM) radio band is a portion of the radio spectrum reserved internationally for the use of radio frequency (RF) energy for industrial, scientific and medical purposes. The frequencies and power of these bands vary from country to country. The most common frequencies encountered are the 868 MHz band – Europe, the 915 MHz band – North America, South America and some other countries, and the 2.4 GHz band – nearly worldwide. In Europe, the use of the ISM band is covered by Short Range Device regulations issued by the European Commission. In the United States, uses of the ISM bands are governed by the Federal Communications Commission (FCC). Short Range Devices (SRD) is a general term, applied to various radio devices designed to operate on a licence-exempt basis, over short range and at low power levels. This includes alarms, telemetry and tele-command devices with maximum power of up to 500 mW at VHF/UHF (very high frequency/ultra high frequency). Frequencies are made available for licence-free operation in the bands of 433 MHz and 868 MHz in Europe and 915 MHz in the USA, Australia and Canada. Modules on different frequencies have different performance characteristics. The 433 MHz frequency has good penetration through structures but low transmit power (10 mW) and hence a limited range. The higher frequency 868 MHz gives improved transmit range thanks to 500 mW power, and can provide 10 km line of sight, and up to 1.3 km transmission lengths through urban environments. It must be noted that both radio frequencies are lower than those used by 2G and 3G services, so they perform better in below ground structures but have far less range than cellular systems due to the limited power. The ISM frequencies are used in networks often referred to as low power wide area networks (LPWAN) used by WiFi (802.11), LoRa, Bluetooth and ZigBee devices (LoRaWan, 2015). However, recommending a particular technology for a specific field application is difficult as the different technologies are developing rapidly in a highly competitive market and have specific relative strengths and weaknesses. This frequency band has a good range if unobstructed, but limited penetration through structures. One of the most interesting aspects of these technologies is the straightforward ability to form local networks as their transmission range in above ground urban areas is a few hundred metres, but their range can be extended indefinitely by organizing them into mesh networks (See *et al.*, 2012). The disadvantage of this type of network is the cost of the additional communication/repeater hubs required to obtain the desired transmission range. For underground applications there is significant attenuation loss and transmission ranges are consequently reduced. The key aspects for this type of mesh-based networks are required range, battery life, robustness to interference and data transfer capacity. Such communication networks have begun to be deployed for research studies associated with urban drainage systems. Ebi *et al*. (2019) reported on the development of a LoRaWAN based arrangement of underground sensor nodes, synchronized to close by over ground repeater nodes which can then transmit data to more remote gateways. At two sites with test periods up to two months, they demonstrated high reliability of data transfer, although data transfer rates were very limited. These studies have demonstrated the flexibility, low cost and adaptability of such radio telemetry-based systems.

5.3 FROM DATA FILES TO STRUCTURED DATABASE

5.3.1 Principles and advantages of relational databases

Data are usually retrieved from the field as unstructured files, either text or binary files. They can be stored as such on a personal computer, or in a shared repository of the organization or company intranet, if they are to be shared between several users. It is common that data are imported in Microsoft Excel: it is a widely available software that most know how to use. Embedded calculation formulas, programming functionalities and automated graph generation allow the user to get a first insight into the data without too much effort. However, these solutions, even if they are certainly widely used, present major drawbacks:

- Apart from very small datasets, the data have to be structured manually: the files are organized in several folders, and/or in worksheets in the case of a workbook. One must be very rigorous to keep the structure as the dataset grows and becomes more complex. Eventually it can become impossible to maintain.
- In some cases, the data files do not contain all the information required to use the data, such as variable names, units, time zones, contact of data producer, types of sensors and so on. These metadata must be added to the dataset, thus adding complexity for manual maintenance, particularly when several people are involved in data collection.
- Access to the data themselves is difficult and slow: one has to find the right file in the right folder, open it or import the data into a processing software without making errors with the format and so on.
- There is no possibility of sharing the data while ensuring its integrity and consistency: copies of the file can be multiplied in several places, files can be modified, updates are not necessarily transmitted.

These limitations can be overcome using relational databases. In a relational database, information is organized in bidimensional tables (or relations, with columns and rows) that are related to each other. Each record in a table (i.e. each row) can be identified individually using a unique identifier (usually an integer value) that is called a primary key. The connection between tables is achieved using foreign keys. It consists of adding a column to a table containing the identifier for each row of the related row in the other table. All complex relations between tables can usually be decomposed into simpler ones.

The main advantages of relational databases compared to unstructured files are:

- Information is not often duplicated, thanks to the use of primary and foreign keys. Thus, there is no information redundancy. The data volume is lower, which optimizes storage.
- The structure of the database is fixed at once, which ensures a better consistency of data and a better sustainability of the database, as the dataset grows over the years. There can be constraints on the data (e.g. one value per date for a given time series, to ensure that there are no duplications).
- Relational databases use a specific programming language for accessing the data: SQL (Structured Query Language). It is a standardized language that allows quick and efficient access to the data, whatever the size of the dataset, which can be unlimited. Complex manipulations of data can also be carried out with SQL.

All these functionalities are implemented in dedicated software tools called Data Base Management Systems (DBMS). Most of these tools come also with additional functionalities such as:

- Data centralization (on a server) and management of access rights (data integrity/security).
- Automatic backups.
- Spatial extension/compatibility with GIS software tools.

5.3.2 Existing DBMS and software solutions

Several Data Base Management Systems are available on the market. A short list of the five 'best known' DBMS is presented in Table 5.1. What distinguishes the software tools are mostly the licence conditions (proprietary vs. free and open-source), the maximum size allowed for the database, and the possibility of adding spatial data into the database. This spatial extension feature makes databases compatible with GIS (geographical information system) software, such as QGIS or ArcGIS. GIS software can use a database with a spatial extension as a centralized data source. The database can then benefit from the functionalities of the GIS software (data visualization, maps and so on). Microsoft Access is distributed with the Microsoft Office Suite, and as such it is the first DBMS that one may come across. It is suitable for small datasets (due to the size limitations) that do not need to be shared between several people. It has no server service. This means that each database is one file stored on a personal computer with no access control: it can be duplicated or modified by any user of the computer. MySQL is the backbone behind most of the commercial websites in the world. PostgreSQL and OracleDB are more specialized in scientific applications. The main difference between them is the licence conditions. PostgreSQL is the open-source reference for structuring and management of environmental observations. It is used for the

Table 5.1 Main database management systems and their key characteristics.

	Max size of a table	Spatial extension (OGC standards)	Server service	Licence
Microsoft Access	2 Gb/database	No	No	Proprietary
Microsoft SQL Server	2 Gb/table	Yes (partial)	Yes	Proprietary
MySQL	Unlimited	Yes (partial)	Yes	Depends
PostgreSQL	Unlimited	Yes (PostGIS)	Yes	Free and open source
OracleDB	Unlimited	Yes (Oracle Spatial)	Yes	Proprietary

OGC – Open Geospatial Consortium.

storage of data from many research observatories throughout the world, as for example in BDOH that stands for Base de Données pour les Observatoires en Hydrologie (i.e. Database for Hydrologic Observatories, Branger *et al.*, 2014), AMMA-CATCH (Galle *et al.*, 2018), ILMS (Zander *et al.*, 2013), Hydroshare (Heard *et al.*, 2014; Yi *et al.*, 2018) or Drought Observatory (Magno *et al.*, 2018).

All the DBMS (even those with a server service) can be installed on a regular PC (which then becomes a 'local server'). One can then benefit from the robust data structure that they offer, the unlimited size of datasets and the efficiency of the SQL queries for data manipulation and access. However, these systems show their full potential (in particular data sharing with access control, proper handling of multiple users simultaneously and data backups) when installed on a proper server independent from the PCs of the data collectors and users.

Data users can have access to the database through specific applications installed on their own computers (heavy or rich client), or more simply with a web browser (thin client). A typical architecture is called 'three-tier' and involves the data server (that contains the DBMS and the datasets), a dedicated web application (program that accesses the server), and a web server on which the users can connect with their computer browser.

Commercial software dedicated to environmental monitoring and based on DBMS can be found on the market. Aquarius by Aquatic Informatics (https://aquaticinformatics.com/products/) and WISKI by Kusters (https://www.kisters.net/NA/products/wiski/) can be cited as examples. Both rely on Oracle DB databases. These software products provide solutions for data structuration, backups, secured access plus a wide variety of functionalities for data visualization and processing. There are also a few free and open-source alternatives made available for a wider public, such as BDOH and Hydroshare (see Section 5.5) which were initially developed for environmental research applications.

5.3.3 Typical data structuration for environmental time series

As an alternative to pre-packaged commercial or open-source solutions, it is also possible to develop custom-made databases. This requires more time and human resources for the informatic development, but it ensures that the database will be adapted to the specificities of the data and to the needs of the end-users. CUAHSI (Consortium of Universities for the Advancement for Hydrologic Science, a non-profit organization that includes many American universities) proposes guidelines about important aspects for database development, such as defining a controlled vocabulary for the names of the variables (Horsburgh *et al.*, 2014) and a database structure that can be used as a template (Horsburgh *et al.*, 2016). Since most data collected are numerical time series, they must be also supported by metadata, i.e. additional information about the measurement attributes (names, time zones, locations of sensors, physical conditions at the measurement sites, etc.). These metadata are essential for the interpretation and further use of the data. Chapters 7 and 9 highlight the type of metadata to be collected and their importance for data quality assurance, respectively.

An important step when building a database is the design of the data structure, i.e. how to dispatch the information into several tables, and how these tables are related to each other. Several options are possible, according to the specific nature of the data and the objectives of the data collectors and users. For example, if there are several measurement points scattered over a large territory, it will be particularly important to be able to locate these measurement points. In other cases, emphasis can be put on the sensors that produce the data, keeping track of information such as the nature of sensor, brand, series number, and maintenance history (sensor replacement, etc.). In other cases, the database will have to be able to deal with several possible time steps for the time series (or even variable time steps).

For example, in the BDOH database (Branger *et al.*, 2014 – see also Section 5.5.1), one of the main requirements was to be able to manage time series with variable time steps. Another important feature was that each value should be associated with an estimation of data quality and possibly a confidence interval corresponding to the data uncertainty (see Chapter 6 for uncertainty assessment). The database was designed with one single table containing all the measured values (by end of 2020, it contained over 60 million entries), that is represented in Figure 5.1. Each entry has a unique index value (id), a date-time value given at 1 millisecond discretization level, the measured value ('valeur' in Figure 5.1) itself, the min/max values of the uncertainty interval and a code corresponding to the data quality. This data quality code is linked with another table that contains all the possible quality codes for the data, hence the foreign key. There is also a reference to the id of the time series the measured values belong to ('chronique_id', linked to the table 'chronique', e.g. time series). All the metadata (description of the quantities, units, etc.) are stored in the 'chronique' table, which is linked to a table 'station' with monitoring stations localized with their XYZ coordinates (Figure 5.2).

5.3.4 Supply of information to databases

In addition to choosing the appropriate tools and building the database, one crucial point is also to define a processing workflow that ensures that the data actually get into the database and are updated regularly. It can be a fully automated process from *in situ* sensors to the database server and even to the data online publication, or it can include several manual steps such as data collection, data validation, uncertainty assessment, etc. All of these aspects must be taken into account and organized so that no data are lost and the information system is efficient (Horsburgh *et al.*, 2011). One should ensure that the time between actual measurements and the proper storage of the corresponding measured values in the database must not be too long. A delay of 3 years between data collection and storage is not acceptable, although it is unfortunately not that uncommon. In such a case, the memory of data acquisition, data validation and data processing is lost and leads to poor final quality.

5.4 DATABASE INTEROPERABILITY

5.4.1 Definition and interest

Interoperability can be defined as a set of characteristics of a system, that allows it to exchange information seamlessly with other systems. In the world of data management, interoperability means that several databases can communicate with each other and exchange information on the data they contain (metadata) and/or data itself. It is one of the pillars of the FAIR data concept.

Interoperability presents huge advantages for data management and exploitation. Instead of having to build and maintain one large and centralized database that must contain everything, it allows entities to build several smaller and specialized databases, that are easier to design and maintain, especially if the data they contain are produced by different entities. It can also bring together data of different sorts that come from various sources, thus providing additional elements for data interpretations. A typical example could be bringing together maps of the sewer system and stormwater control measures (SCMs), logs of the operations on weirs, and continuous flow measurements from *in situ* sensors. Mostly these three types of data are not available together, at least not in a straightforward way. One person has to request them from two or three different people, each time a question arises. A change of direction of flow monitored by the sensors could typically be interpreted after several weeks of investigation as due to an operation on a weir in a nearby branch of the sewer. If these data were stored in interoperable systems, they would be known to all the systems (and operators) as soon as they are uploaded in their respective databases, thus saving a lot of time and energy. Finally, interoperability is the cornerstone of public data portals, such as Data Grand Lyon (see

id	qualite_id	chronique_id	date	valeur	minimum	maximum
11076435	1	150	2006-02-12 13:30:00	7	*NULL*	*NULL*
11076436	1	150	2006-02-12 14:00:00	6.98	*NULL*	*NULL*
11076437	1	150	2006-02-12 14:30:00	6.98	*NULL*	*NULL*
11076438	1	150	2006-02-12 15:00:00	6.98	*NULL*	*NULL*
11076439	1	150	2006-02-12 15:30:00	6.98	*NULL*	*NULL*
11076440	1	150	2006-02-12 16:00:00	6.98	*NULL*	*NULL*
11076441	1	150	2006-02-12 16:30:00	6.98	*NULL*	*NULL*
11076442	1	150	2006-02-12 17:00:00	6.97	*NULL*	*NULL*
11076443	1	150	2006-02-12 17:30:00	6.97	*NULL*	*NULL*
11076444	1	150	2006-02-12 18:00:00	6.95	*NULL*	*NULL*
11076445	1	150	2006-02-12 18:30:00	6.97	*NULL*	*NULL*
11076446	1	150	2006-02-12 19:00:00	6.95	*NULL*	*NULL*
11076447	1	150	2006-02-12 19:30:00	6.97	*NULL*	*NULL*
11076448	1	150	2006-02-12 20:00:00	6.95	*NULL*	*NULL*
11076449	1	150	2006-02-12 20:30:00	6.95	*NULL*	*NULL*
11076450	1	150	2006-02-12 21:00:00	6.95	*NULL*	*NULL*
11076451	1	150	2006-02-12 21:30:00	6.93	*NULL*	*NULL*
11076452	1	150	2006-02-12 22:00:00	6.95	*NULL*	*NULL*
11076453	1	150	2006-02-12 22:30:00	6.93	*NULL*	*NULL*
11076454	1	150	2006-02-12 23:00:00	6.92	*NULL*	*NULL*
11076455	1	150	2006-02-12 23:30:00	6.92	*NULL*	*NULL*
11076456	1	150	2006-02-13 00:00:00	6.93	*NULL*	*NULL*
11076457	1	150	2006-02-13 00:30:00	6.93	*NULL*	*NULL*
11076458	1	150	2006-02-13 01:00:00	6.92	*NULL*	*NULL*
11076459	1	150	2006-02-13 01:30:00	6.91	*NULL*	*NULL*
11076460	1	150	2006-02-13 02:00:00	6.91	*NULL*	*NULL*
11076461	1	150	2006-02-13 02:30:00	6.91	*NULL*	*NULL*
11076462	1	150	2006-02-13 03:00:00	6.89	*NULL*	*NULL*
11076463	1	150	2006-02-13 03:30:00	6.91	*NULL*	*NULL*
11076464	1	150	2006-02-13 04:00:00	6.9	*NULL*	*NULL*

Figure 5.1 Extract of the main storage table in BDOH. *Source*: Flora Branger (INRAE) (Screenshot of the BDOH database).

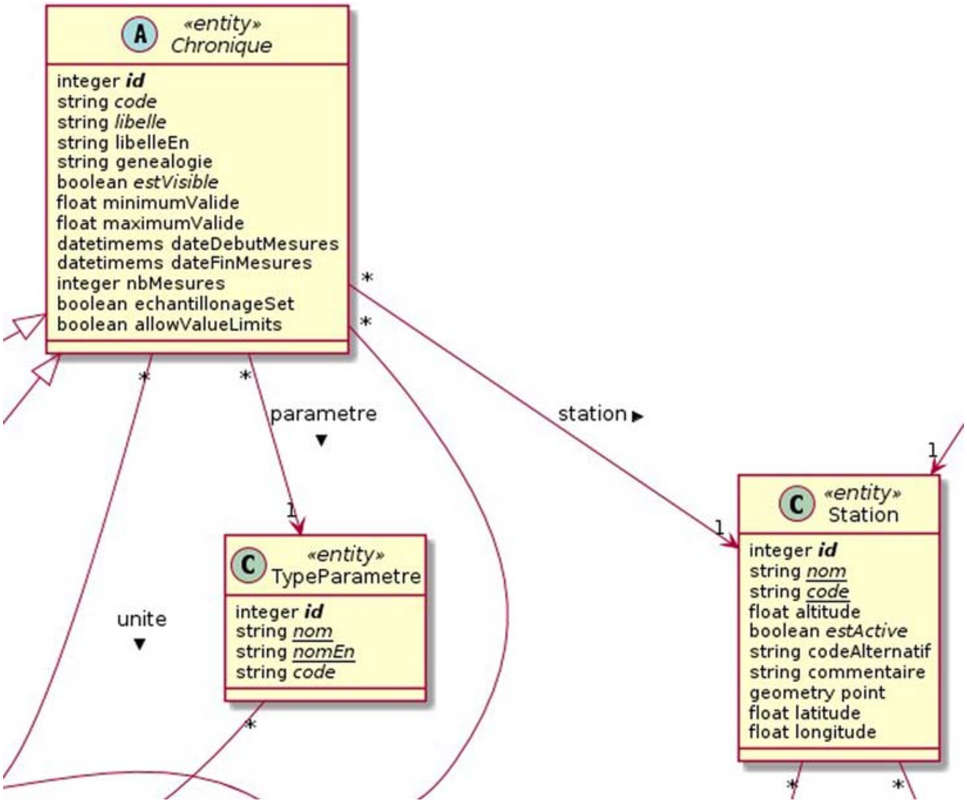


Figure 5.2 Extract of BDOH database structure: chronique and station tables. *Source*: Flora Branger (INRAE) (Screenshot of the BDOH database).

Section 5.5.3) or the Theia/OZCAR Information System (Braud *et al.*, 2020). Such data portals present and give access to datasets that come from various data producers and are stored in various databases. They are the full implementation of the FAIR data concept, and as such, can be considered as the future of data storage.

5.4.2 Interoperability standards and examples

As for sensors, interoperability means that communication protocols must be defined for databases. These communication protocols operate through the web, and as such are called webservices. In the field of environmental data, the Open Geospatial Consortium (OGC) is in charge of defining open standards for these webservices. The OGC is an international standards organization, that was created in 1994, initially to foster the development of open geospatial tools. The most well-known and widely used standard webservices proposed by the OGC are the CSW (Catalogue Service for the Web) that is behind online metadata catalogues (such as https://www.data.gouv.fr/fr/ (accessed 07 Dec 2020) in France, https://www.ukdataservice.ac.uk/ (accessed 07 Dec 2020) in the UK, etc.) and information sharing between them (called metadata harvesting: information added in one catalogue is automatically replicated in the others), and also the WMS (Web Map Service) and WFS (Web Feature Service) that are behind most displays of interactive maps on the web.

In terms of *in situ* measurements, several standards are proposed by the OGC:

- The Sensor ML standard (https://www.ogc.org/standards/sensorml (accessed 07 Dec 2020)) describes the measurement process of observation data (description of sensors and also post-measurement data processing) but does not contain the measured values themselves.
- The Observation & Measurement (O&M) standard (http://www.opengeospatial.org/standards/om (accessed 07 Dec 2020)) describes the properties of observations, including surrounding environment features (location, sampling strategy if any, organization). It is complementary to the SensorML standard and can contain the actual data.
- The Sensor Observation Service (SOS) (http://www.opengeospatial.org/standards/sos (accessed 07 Dec 2020)) defines a webservice interface for querying real-time sensor data and sensor metadata. It uses the information encoded in SensorML (description of Sensor) and O&M (description of measurements).

These three standards are the key to the implementation of data operability. They are not specialized to water-related data, and are generic enough to handle any kind of observation data. More specific to hydrological data, the WaterML standard (https://www.ogc.org/standards/waterml (accessed 07 Dec 2020)) has been defined more recently. It introduces additional concepts specific to hydrological time series, such as gauging and rating curves, or temporal data interpolation. However, it is not as widely used as the three other standards presented here.

OGC provides standards and specifications, but does not provide ready-to-use tools. The implementation of the standards has to be carried out. The standards are long and complex, they do not have to be implemented to their full extent. Several implementations of the SOS webservice are available. The most widely used implementation, recognized as reference, is the open-source German-based 52°North solution (https://52north.org/software/software-projects/sos/ (accessed 07 Dec 2020)). This solution also takes into account the WaterML standard. The 52°North SOS is currently behind many environment data portals, and in particular behind the case study Data Grand Lyon.

5.4.3 Practical recommendations

Data portals that implement interoperability are high level software constructions. As a water scientist or a practitioner, the objective is not to be able to implement them or manage them directly, but just to understand the underlying concepts enough to be able to communicate with the specialists. The development and maintenance of interoperability in databases requires skilled software developers and IT support. However, many organizations now do have the corresponding human resources, in universities or research institutes, and in local authorities' administrations (regional councils, city councils).

Although we can see interoperable data portals as the future of data storage, we must also keep in mind that the data must be properly structured and stored beforehand, following the guidelines given in the previous sections. If the data is not structured and available online, there is no way it can become interoperable.

5.5 CASE STUDIES

Four case studies are presented to illustrate emerging approaches to: (i) local wireless-based communication, and (ii) the organization of collected urban drainage system field data.

5.5.1 Case study 1: BDOH (Base de Données des Observatoires en Hydrologie), a database for the storage and publication of long-term water observation data

5.5.1.1 Context and objectives

BDOH has been developed by Irstea (now INRAE) in France, which has been responsible since 2013 for storage, management and dissemination of hydrological time series produced in its long-term environmental observatories (Figure 5.3). The main recorded quantities are rainfall, streamflow, groundwater level, soil moisture, air and water temperature, suspended solids concentration as well as concentrations of various chemical substances. At the end of 2020, BDOH contained records from 12 observatories at different locations in France, and over 55 years, which makes approximately 60 million records (https://bdoh.irstea.fr (accessed 07 Dec 2020)). Data can be browsed and visualized freely. Data download is free upon authentication for all users (scientists, management authorities, general public). BDOH now has almost 300 registered users.

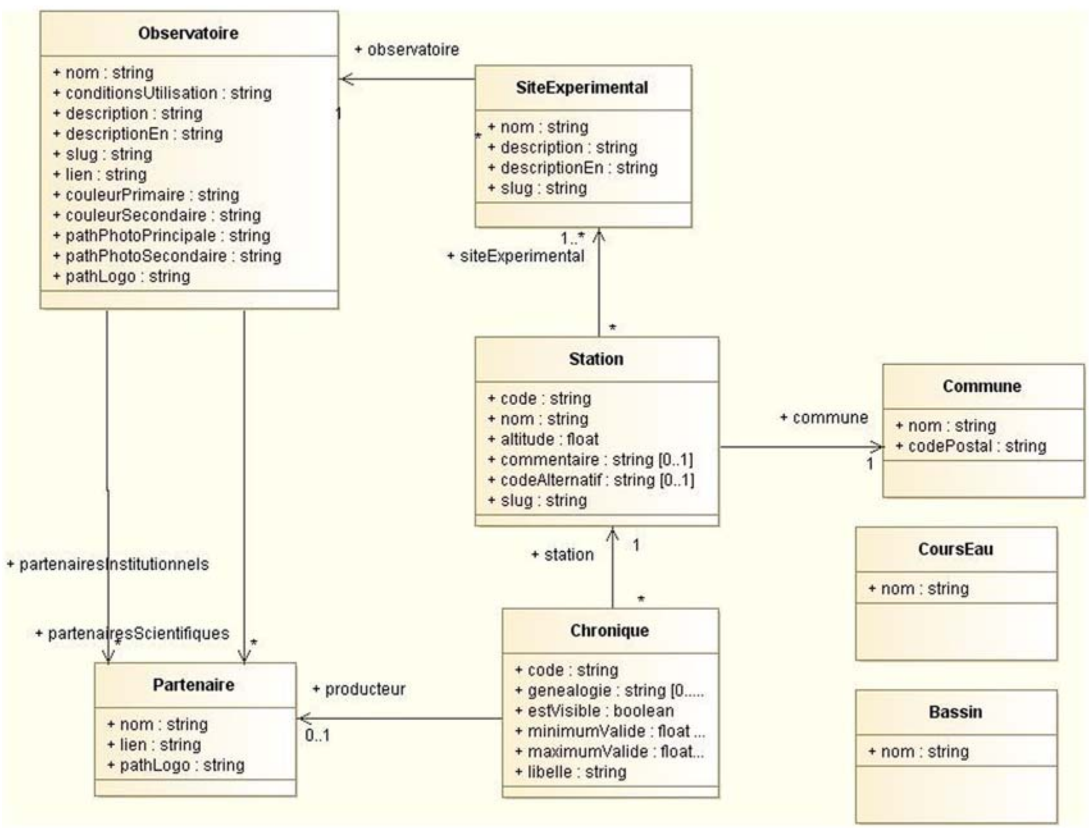


Figure 5.3 General structure of the BDOH database. *Source*: Flora Branger (INRAE) (Screenshot of the BDOH database).

5.5.1.2 Structure and main features

BDOH is based on open-source technologies such as the database management system PostgreSQL/PostGIS, and a web application developed in PHP and Javascript. The structure of the database considers stations that are points in space where physical quantities are being measured. The time series for each of these quantities are stored at variable time steps. Each value is associated with a date/time, a code appreciating the quality of the data (see also Section 9.2.2), and an estimation of the measurement uncertainty (see Section 8.4). Functionalities for data managers include data import from flat files (text files) in various formats, automatic calculation of time series derived from other time series (for example streamflow calculation from water level measurements using rating curves, and taking into account rating curve changes over time, or contaminant fluxes, see Section 9.4.3.3), and import and visualization of manual control points. An important point is that each data must be associated with a specific quality code, which means that data must be critically reviewed/validated before being imported into BDOH. Additional functionalities are also available for data users, such as search engine, graphical visualization (Figure 5.4), and options for data download (file formats, time steps).

5.5.1.3 Advantages and drawbacks

The use of a relational database management system and a web application means that the data is easy to browse and to extract, and is permanently available. It is also hosted on a server with automatic backups and security checks so that the integrity of the data can be ensured. The qualification of the data is also

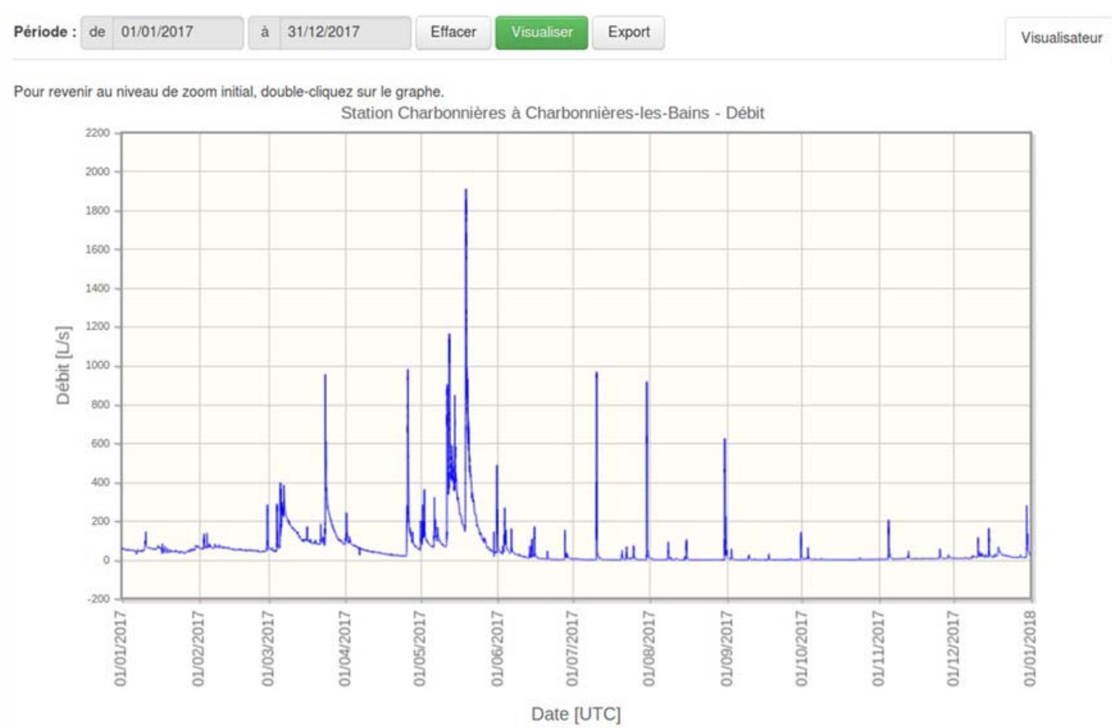


Figure 5.4 Snapshot of data visualization in BDOH. *Source*: Flora Branger (INRAE) (Screenshot of the BDOH database https://bdoh.irstea.fr/YZERON/V3015301/DEB (accessed 07 Dec 2020)).

an asset, although BDOH does not provide a method or workflow on how to assess this quality. BDOH also fostered exchanges and general improvement regarding data production practices. The main drawback is that of course it is quite some work to install and set up a BDOH instance. It requires computer resources and skilled human resources, in particular an IT system that can manage a Linux server. Otherwise, the time required for maintenance and data administration is quite low, approximately a 10% part-time job for over 60 million records.

5.5.2 Case study 2: DoMinEau, an Excel-based database for water quality monitoring

5.5.2.1 Context and objectives

The DoMinEau database (http://www.graie.org/Sipibel/projets.html#DOMINEAU (accessed 07 Dec 2020)) was developed in 2016 in response to a call by Agence Française de la Biodiversité (now Office Français de la Biodiversié, a French public national agency in charge of fostering research and expertise on biodiversity as well as helping to apply public water policies) and for the needs of the SIPIBEL observatory (Site PIlote de BELlecombe – Bellecombe Pilot Site). SIPIBEL investigates pollutants in urban and hospital wastewater and fosters research programmes in four themes (pollutant flows, treatment, risks, and sociology). The objectives of DoMinEau were to structure the datasets to (i) provide information about the pollution of wastewater and receiving water, including micropollutants, microbiology and bio-indicators, (ii) provide an estimation of the data quality and (iii) make the datasets easily accessible, searchable and interoperable by the partners of the project (scientists and water management authorities). At the end of the project, the DoMinEau database contained over 55,000 records.

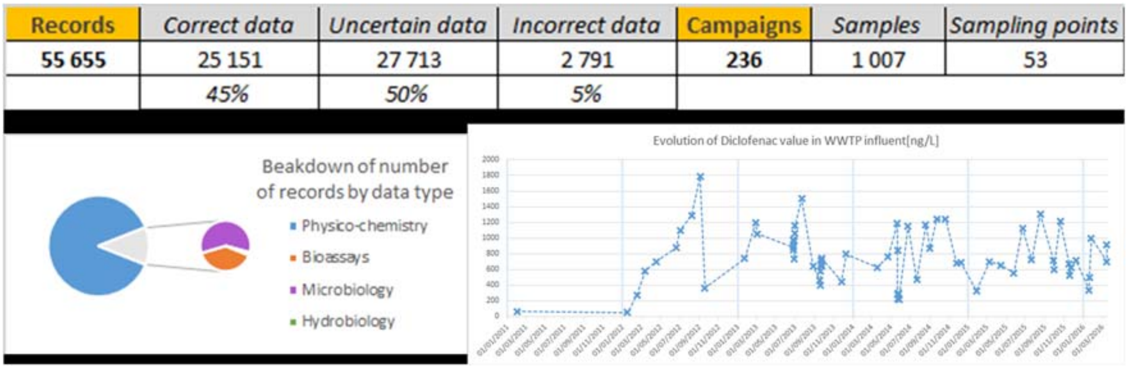
5.5.2.2 Structure and main features

DoMinEau is based on Microsoft Excel. This software was chosen because of its widespread availability, because all the targeted users (scientists and water management authorities) were familiar with its use, and because there was little time awarded in the project for database development.

DoMinEau consists of four workbooks, each one organized with several worksheets:

- Sites-parameters-and-methods: contains all the metadata associated with the pollutant datasets, incl. sampling points, quantities, analytical methods. The list of quantities includes physical-chemical (temperature, water flow, pH, conductivity), micropollutants, microbiology, hydrobiology indicators and bio-essay indicators. There can be several analytical methods for each quantity.
- Campaigns-and-results: contains the sampling dates, sample identifiers, analyses results and a quality assessment for each individual result. The structure is flexible enough to record any type of sample at any frequency that was collected by the observatory. A sample is defined by a sampling point (location), the start and end dates of sampling and its duration, the type of sample (flow dependent or not) and the conditions during the sampling (river discharge, 48-hrs rainfall, etc.). The quality assessment method is shown in Figure 5.6.
- Statistics-and-graphics: calculates and displays automatically summary statistics and graphs about the data (see Figure 5.5).
- Data extraction: generates files that can be loaded into statistics and data analysis software, such as R or Matlab®. This is only an extraction utility; it does not carry out additional calculations.

A complete workflow was developed for quality assessment. Seven indicators (Figure 5.6) are used to assign a quality category (correct, dubious, incorrect) to each data. It takes into account not only the



Records	Correct data	Uncertain data	Incorrect data	Campaigns	Samples	Sampling points
55 655	25 151	27 713	2 791	236	1 007	53
	45%	50%	5%			

Figure 5.5 Snapshot of summary metadata statistics from the DoMinEau database. *Source*: GRAIE (http://www.graie.org/Sipibel/projets.html – DOMINEAU (accessed 07 Dec 2020)).

quality of the chemical analysis, but also the quality of the sampling itself. For each indicator there is a drop-down list with prepared values (pre-calculated). These values are set manually by the person who imports the data into the database. This step introduces some subjectivity, the reader is referred to the discussion on this important aspect in Chapter 9 on data validation. The most critical step is the coordination with the persons who were actually in the field and collected the samples. At the beginning of the DoMinEau database this was the same person as the data administrator, so that was quite easy. At present, the data administrator works with copies of the field sheets and contacts the data providers if necessary.

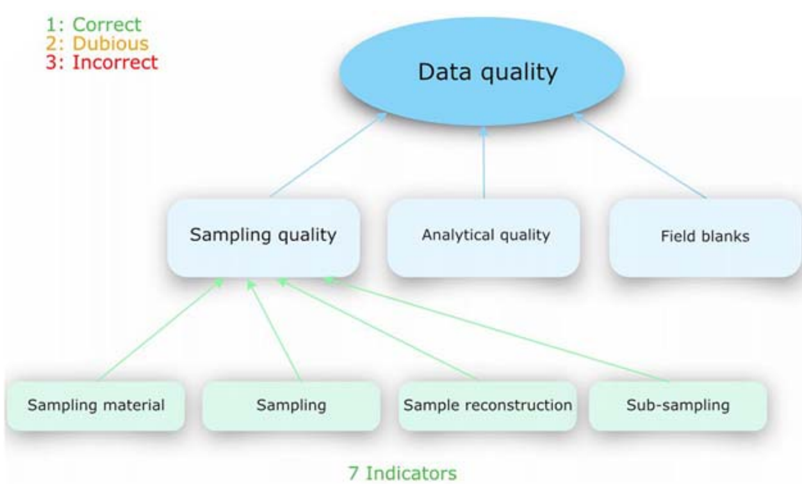


Figure 5.6 Criteria for data quality assessment in DoMinEau. For each indicator, three levels are possible. The overall data quality is assessed by the count of indicators in each colour. Overall quality is green if all indicators are green. Only one orange indicator triggers an overall orange quality. If there is one red value or more than three orange values the overall quality is red. *Source*: Flora Branger (INRAE).

DoMinEau is managed by a data administrator at GRAIE (a non-profit organization in charge of the coordination of the SIPIBEL observatory and of data management). In order to disseminate the data to the partners, copies of the Excel database are sent over on request. In addition, the data were all uploaded to the NORMAN network database system (European network for monitoring of emerging environmental substances (https://www.norman-network.com/nds/ (accessed 07 Dec 2020)).

5.5.2.3 Advantages and drawbacks

The main advantage of the DoMinEau database is the effort put on the structure and validation of the data, which results in a more rigorous approach for field work and in particular for compliance with the sampling protocols. It has contributed greatly to improve the quality of the data, which is a good example of a 'virtuous circle'. The statistics and graphic modules provide a first quick overview of the data in real time (as soon as the data is uploaded in the database) that is useful to compare campaigns. The summary dashboard is also useful for project managers and for achieving the project deliverables.

The main drawback is caused by the limitations of the selected software, i.e. Microsoft Excel. In terms of data management, new data must be added manually as new lines to the spreadsheet. With 55,000 records, this is quite cumbersome and it is very easy to make mistakes. The macros (in particular for the automatic statistics and graph generation) can be slow. The system is sensitive to successive updates and versions of Excel. Data management thus requires a lot of time from the Data Administrator (50% FTE [full time equivalent] approximately) and the database being off-line (the only way to transfer data is to send over copies), hence the data are not easily accessible. In practice DoMinEau has not been used as expected by the targeted users (scientists and water management authorities) and did not foster data sharing between project partners as expected. Its use requires significant effort and time. The data are now also stored in the online NORMAN database but they are quite difficult to find and mixed with other datasets. In conclusion, although DoMinEau proved to be useful in several ways during the SIPIBEL project, the technical choice of Excel appeared to limit its impact and the dissemination of the data during and after the project. It shows that alternative and more sustainable storage techniques would be preferable for future projects.

5.5.3 Case study 3: Data Grand Lyon – open data portal

The platform Data Grand Lyon (https://data.grandlyon.com/accueil (accessed 07 Dec 2020)) is far more ambitious than just a water data portal. It has been under development since 2011 as part of the Smart City policy of the Greater Lyon metropolis. Its ambition is to make available all types of data produced over the metropolis territory, that can be used:

- To facilitate communication and information exchange between the various management authorities.
- To encourage citizens to access the data and encourage potential participation.
- By making data freely available to foster innovation and economic initiatives in terms of data analysis and visualization.

These data contain items such as aerial photographs, street and cadastral maps, state of traffic, noise data, air temperature data, availability of Velo'v bikes (local bicycle sharing system), localization of disabled persons parking places, bus network alerts, some being available in real-time. They are produced by different stakeholders over the metropolis territory, public or private, with the ambition of data mutualization and interoperability.

Data are distributed according to three types of licence: open data, data delivered for free upon authentication, and data accessible with a licence fee. Most data are open.

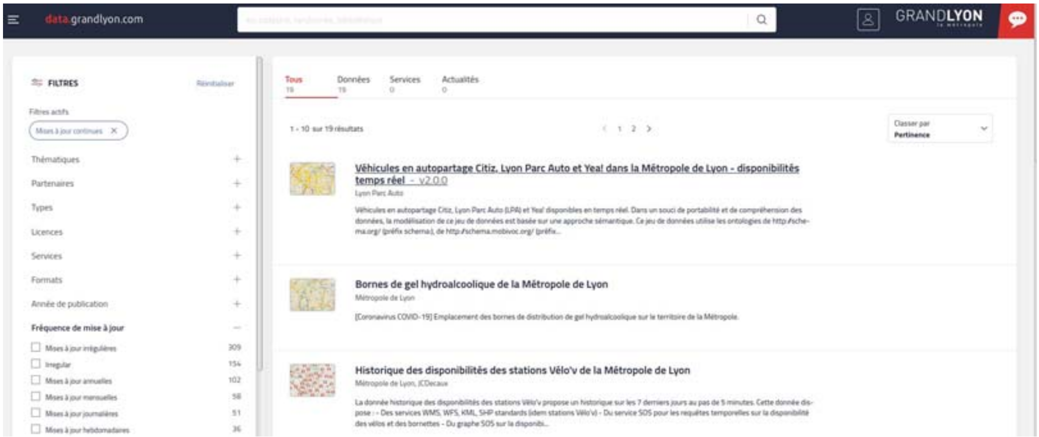


Figure 5.7 Snapshot of Data Grand Lyon metadata catalogue. *Source*: Flora Branger (INRAE) (Screenshot from Data Grand Lyon https://data.grandlyon.com/recherche (accessed 07 Dec 2020)).

Data Grand Lyon is based on a metadata catalogue (Figure 5.7) with a search engine, along with several protocols that:

- Take the data from various sources (relational databases, but also plain files).
- Format them (middle office, Figure 5.8).
- Make them available using (mostly) standard webservices (front office, Figure 5.8).

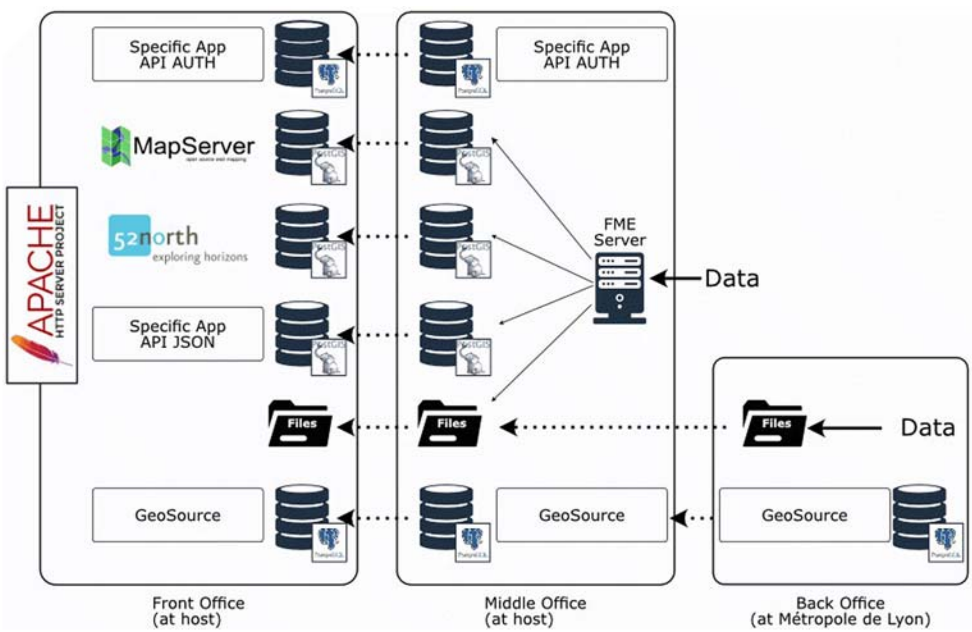


Figure 5.8 Architecture of Data Grand Lyon platform. *Source*: Grand Lyon.

Data Grand Lyon does not provide the basic structure of data such as presented in the BDOH case study (see Section 5.5.1). However, it can deal with unstructured data to some extent (plain files with stabilized formats and protocols for data updates). Data Grand Lyon relies mostly on open-source software. In terms of standards for data exchange (see Section 5.4.2), Data Grand Lyon implements the WMS/WFS standards for geographic data through MapServer. It also implements the SOS (Sensor Observation Service) for time series using the 52N open-source solution.

At the end of 2020, hydrological data were not present in the platform. Only rainfall data (from rain gauges located all over the metropolis territory) are available. They are currently being pushed on to the platform from Excel files. However, hydrological data could seamlessly fit in, using the SOS webservice that can take into account all kinds of data. Data Grand Lyon is an example of the very powerful tools that are now being developed for data storage and access. It is probably the best way to communicate data, even in real-time. However, one must keep in mind that it does not provide the basic structure of data (although it can deal with unstructured data to some extent). And it is also quite expensive to maintain as it requires specific computer resources and skilled data scientists.

5.5.4 Case study 4: local wireless based system for flood risk assessment and reduction – CENTAUR

CENTAUR (Cost-Effective Neural Technique to Alleviate Urban flood Risk, https://www.sheffield.ac.uk/centaur/home/outputs (accessed 16 Dec. 2020)) is an intelligent autonomous system for local urban flood risk reduction. It utilizes underutilized storage capacity to reduce the flow rate and water level at vulnerable potential flood locations. A gate controls flow using an intelligent algorithm which analyses local water level data, and then instructs the gate to adjust to the contemporary conditions. As this is a data driven system, reliable data transfer between the water level sensors and the gate controller is essential. CENTAUR is self-managing and easily deployed as a retrofit solution. It is far less disruptive and significantly less costly than alternative capital and space intensive solutions.

Data transfer for the system is wireless based and the communication equipment is modular and extensible; modules are lamp-post mounted or in-manhole. The types of modules are:

- Monitoring STation (MST): provides real-time monitoring of water level via radio communication between modules; it is ATEX certified for use in sewers with potentially explosive atmospheres.
- Control STation (CST): provides real time control of PLC (programmable logic controller) or analogue signal-controlled flow control devices.
- HUB (HUB): executes a fuzzy logic control algorithm based on the time-series of water level information relayed from MSTs, and relays commands to CSTs; it also communicates periodically with a web-based online dashboard, with system diagnostics transferred intermittently via GSM communications.
- RePeaTer (RPT): provides an extension of radio communication links between modules, when distances between modules are large or line-of-sight is particularly disrupted by obstacles.
- online dashboard: cloud-based dashboard which communicates with the HUB to give remote viewing of system operation, allows remote reconfiguration of modules, and flags any faults (however, the system operates autonomously and independently of the dashboard; the dashboard is in place for operational monitoring).

The technical innovation is around deploying artificial intelligence, autonomy, and tailoring different communication technologies for reliability and power efficiency. The system design exploits mainstream technologies to minimize cost and ensure reliability.

Radio communication for monitoring and issuing control instructions gives guaranteed signal over the typical distances between flow control gates and water level monitoring locations (up to several hundred metres), that characterize upstream storage availability and downstream flood risk locations. Distances can be extended indefinitely by use of a series of RPTs. The radio protocol was specially engineered for the monitoring paradigm; bi-directional, and communicating from below to above ground, and then relaying above ground. During the development of CENTAUR, the off-the-shelf protocol LoRa was unable to provide the reliability required by the system. Although line-of-sight performance was good, obstacles in built up urban areas had a significant detrimental effect consistent with noted reliability of around 90% in Ebi *et al*. (2019). This was insufficient for the CENTAUR application which featured mission-critical control as well as monitoring. An application-specific radio platform was developed based on a chip set from a major international manufacturer. This and the use of repeaters led to the 100% reliability required by the control application. Use of the GSM network was ruled out early in the design process, due to its lack of reliability and its latency. GSM communications are often intermittent and are unsuitable for mission-critical control systems. It is suitable for monitoring applications where missing data or latent communication are acceptable.

The modules can be configured using USB connection or via Bluetooth and a smartphone app. Basic reconfiguration is also achievable via the online dashboard. Bluetooth is particularly convenient for field engineers when installing the system. LED-based diagnostics on the front of each module make for convenient communication establishment and error diagnosis between modules. Such capability is important for applications in which long term, reliable performance is necessary and is often overlooked when data transfer is considered.

In terms of cyber-security, the inter-module proprietary protocol is owned by Environmental Monitoring Solutions Ltd. An attacker would require several key pieces of information to be able to interpret, modify and transmit messages to/from devices. Messages are transmitted in binary form with data embedded anonymously. Use is made of cyclic redundancy checks (CRC) and checksum functions to confirm the integrity of the data; any change to the message would require a change to the CRC and checksums before the receiving device will accept it, even if the attacker could decipher the message content. Additionally, the use of radio communications rather than the mobile phone network and internet steers firmly away from the possibility of cyber-attack. An attacker would need access to specialized listening and transmitting instrumentation rather than just access to the internet.

Communication between the HUB module and the internet-based dashboard is via JSON messages which use SSL security to encrypt messages. CENTAUR servers are virtual machines and are cloud-hosted on a cloud platform. The system therefore leverages the platform provider's cyber security strategy (based on 'assume breach'). A new server and application instance is created for each customer to ensure privacy and security. Bluetooth connections to modules are opened as required, and close automatically on a time-out.

Sensor redundancy gives reliable water level data. The system can disable itself automatically on failure of, for example, a communication link. The gate technology is easily deployed with physical fail-safes to give minimum upstream risk.

The system is self-managing; it is effectively 'fit-and-forget'. CENTAUR first deployment was in Coimbra, Portugal in 2017, where it controlled around 60 storms in its first year of operation delivering a tangible reduction in flood risk. CENTAUR can also be used for limiting CSO spills, managing flows into energy intensive assets using local water level data transferred using the radio-based communication system described above.

5.6 SUMMARY AND TRANSITION

This chapter summarized the main wired and wireless technologies that can be used to transfer data from the measurement location in UDSM systems to the location/database where data can be converted into understanding of a system state. From the raw data to databases that allow communication with other ones, this entire process requires hardware, services and IT competences to be conducted properly.

Chapters 2 to 4 list the most common sensors for monitoring and here Chapter 5 describes how to transfer, record and store data, i.e. the required hardware to transfer and then record data and make them available for any future use. Once those choices have been made, monitoring network and stations must be designed. Chapter 6 gives details on the macro- (network) and micro- (stations) design of monitoring systems whose processes for data-acquisition, -storage and -access are of key importance. There are also strong links between the current chapter and subsequent Chapters 7 to 9 in which operation & management, and data uncertainty and data validation are discussed, respectively. In these chapters the importance of employing a high-quality database will become even more clear. As each aspect of monitoring is essential in contributing to obtain information in a controlled and structured manner, a well-wrought design of the data handling allows reconstruction of the 'history' of the data offered to a range a data users. This includes data on maintenance activities, calibration results, validation results and, depending on the application, processed data.

REFERENCES

Branger F., Thollet F., Crochemore M., Poisbeau M., Raidelet N., Farissier P., Lagouy M., Dramais G., Le Coz J., Guérin A., Tallec G., Peschard J., Mathys N., Klotz S. & Tolsa M. (2014). Le projet Base de Données pour les Observatoires en Hydrologie : un outil pour la bancarisation, la gestion et la mise à disposition des données issues des observatoires hydrologiques de long terme à Irstea [Database for hydrological observatories: a tool for storage, management and access of data produced by the long-term hydrological observatories of Irstea]. *La Houille Blanche*, **1**, 33–38. doi: 10.1051/lhb/2014005.

Braud I., Chaffard V., Coussot C., Galle S., Juen P., Alexandre H., Baillion P., Battais A., Boudevillain B., Branger F., Brissebrat G., Cailletaud R., Cochonneau G., Decoupes R., Desconnets J.-C., Dubreuil A., Fabre J., Gabillard S., Gérard M.-F., Grellet S., Herrmann A., Laarman O., Lajeunesse E., Le Hénaff G., Lobry O., Mauclerc A., Paroissien J.-P., Pierret M.-C., Silvera N. & Squividant H. (2020). Building the information system of the French Critical Zone Observatories network: Theia/OZCAR-IS. *Hydrological Sciences Journal*, in press. doi: 10.1080/02626667.2020.1764568.

Ebi C., Schaltegger F., Rüst A. & Blumensaat F. (2019). Synchronous LoRA mesh network to monitor processes in underground infrastructure. *IEEE Access*, **7**, 577663–557677. doi: 10.1109/ACCESS.2019.2913985.

EU (1991). Council Directive 91/271/EEC of 21 May 1991 concerning urban waste-water treatment. *Official Journal of the European Communities*, 30 May, L135/40-L135/52. Available at https://op.europa.eu/en/publication-detail/-/publication/34860b42-95cc-4a57-b998-2e445d5befa9/language-en (accessed 15 December 2020).

EU (1998). Commission Directive 98/15/EC of 27 February 1998 amending Council Directive 91/271/EEC with respect to certain requirements established in Annex I thereof (Text with EEA relevance). *Official Journal of the European Communities*, 07 March, L67/29-L67/30. Available at https://op.europa.eu/en/publication-detail/-/publication/ff7ec087-8cc3-4619-bffc-b08ea4883d2c (accessed 15 December 2020).

EU (2000). Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for Community action in the field of water policy. *Official Journal of the European Union*, 22 Dec., L327/1-L327/73. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32000L0060&from=EN (accessed 16 December 2020).

EU (2019). Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information. *Official Journal of the European Union*, 26 June, L172/56-L172/83.

Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1024&from=EN (accessed 14 December 2020).

Galle S., Grippa M., Peugeot C., Bouzou Moussa I., Cappelaere B., Demarty J., Mougin E., Panthou G., Adjomayi P., Agbossou E. K., Ba A., Boucher M., Cohard J.-M., Descloitres M., Descroix L., Diawara M., Dossou M., Favreau G., Gangneron F., Gosset M., Hector B., Hiernaux P., Issoufou B.-A., Kergoat L., Lawin E., Lebel T., Legchenko A., Malam Abdou M., Malam-Issa O., Mamadou O., Nazoumou Y., Pellarin T., Quantin G., Sambou B., Seghieri J., Séguis L., Vandervaere J.-P., Vischel T., Vouillamoz J.-M., Zannou A., Afouda S., Alhassane A., Arjounin M., Barral H., Biron R., Cazenave F., Chaffard V., Chazarin J.-P., Guyard H., Koné A., Mainassara I., Mamane A., Oi M., Ouani T., Soumaguel N., Wubda M., Ago E. E., Alle I. C., Allies A., Arpin-Pont F., Awessou B., Cassé C., Charvet G., Dardel C., Depeyre A., Diallo F. B., Do T., Fatras C., Frappart F., Gal L., Gascon T., Gibon F., Guiro I., Ingatan A., Kempf J., Kotchoni D. O. V., Lawson F. M. A., Leauthaud C., Louvet S., Mason E., Nguyen C. C., Perrimond B., Pierre C., Richard A., Robert E., Román-Cascón C., Velluet C. & Wilcox C. (2018). AMMA-CATCH, a critical zone observatory in West Africa monitoring a region in transition. *Vadose Zone Journal*, **17**(1), 1–24. doi: 10.2136/vzj2018.03.0062.

Heard J., Tarboton D., Idaszak R., Horsburgh J., Ames D., Bedig A., Castronova A. M., Couch A., Dash P., Frisby C., Gan T., Goodall J., Jackson S., Livingston S., Maidment D., Martin N., Miles B., Mills S., Sadler J., Valentine D. & Zhao L. (2014). An Architectural Overview of Hydroshare, a Next-generation Hydrologic Information System. *Proceedings of the 11th International Conference on Hydroinformatics, HIC 2014*, New York City, USA, Paper 311, 8 p. http://academicworks.cuny.edu/cc_conf_hic/311/ (accessed 31 May 2021).

Horsburgh J. S., Tarboton D. G., Maidment D. R. & Zaslavsky I. (2011). Components of an environmental observatory information system. *Computers and Geosciences*, **37**(2), 207–218. doi: 10.1016/j.cageo.2010.07.003.

Horsburgh J. S., Tarboton D. G., Hooper R. P. & Zaslavsky I. (2014). Managing a community shared vocabulary for hydrologic observations. *Environmental Modelling and Software*, **52**, 62–73. doi: 10.1016/j.envsoft.2013.10.012.

Horsburgh J. S., Aufdenkampe A. K., Mayorga E., Lehnert K. A., Hsu L., Song L., Jones A. S., Damiano S. G., Tarboton D. G., Valentine D., Zaslavsky I. & Whitenack T. (2016). Observations Data Model 2: a community information model for spatially discrete earth observations. *Environmental Modelling and Software*, **79**, 55–74. doi: 10.1016/j.envsoft.2016.01.010.

LoRaWAN (2015). *LoRaWAN: what is it? A technical overview of LoRA and LoRaWAN*. LoRa Alliance, San Ramon, CA (USA), Technical Marketing Workgroup 1.0, 20 p. Available at https://lora-alliance.org/sites/default/files/2018-04/what-is-lorawan.pdf (accessed 7 December 2020).

Magno R., De Filippis T., Di Giuseppe E., Pasqui M., Rocchi L. & Gozzini B. (2018). Semi-automatic operational service for drought monitoring and forecasting in the Tuscany region. *Geosciences*, **8**(2), 49, 24 p. doi: 10.3390/geosciences8020049.

See C. H., Horoshenkov K. V., Abd-Alhameed R. A., Hu Y. F. & Tait S. J. (2012). A low power wireless sensor network for gully pot monitoring in urban catchments. *IEEE Sensors Journal*, **12**(5), 1545–1553. doi: 10.1109/JSEN.2011.2174354.

Wilkinson M. D., Dumontier M., Aalbersberg I. J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L. B., Bourne P. E., Bouwman J., Brookes A. J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C. T., Finkers R., Gonzalez-Beltran A., Gray A. J. G., Groth P., Goble C., Grethe J. S., Heringa J., 't Hoen P. A. C., Hooft R., Kuhn T., Kok R., Kok J., Lusher S. J., Martone M. E., Mons A., Packer A. L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.-A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M. A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J. & Mons B. (2016). Comment: the FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018, 9 p. doi: 10.1038/sdata.2016.18.

Yi H., Idaszak R., Stealey M., Calloway C., Couch A. L. & Tarboton D. G. (2018). Advancing distributed data management for the HydroShare hydrologic information system. *Environmental Modelling and Software*, **102**, 233–240. doi: 10.1016/j.envsoft.2017.12.008.

Zander F., Kralisch S. & Flügel W.-A. (2013). Data and information management for integrated research – Requirements, experiences and solutions. *Proceedings of the 20th International Congress on Modelling and Simulation, MODSIM 2013*, 1–6 December, Adelaide, Australia, 2201–2206.