Chapter 9

Data handling and storage

D. Prodanović[1]

[1] Institute for Hydraulic and Environmental Engineering, Faculty of Civil Engineering, University of Belgrade, Bulevar Kralja Aleksandra, Belgrade, Serbia

9.1 INTRODUCTION

Integrated urban water management is likely to involve the collection of vast amounts of data, from many different sources. To achieve true integration, the accumulated data must be able to be easily shared among different users. This chapter therefore discusses:

- portability of data formats,
- data application for different models,
- lifetime of data and storage systems, and
- data security.

Firstly, a short historical overview is given, followed by an explanation of typical data flow processes in integrated urban water management (IUWM) systems. Since the database management system is at the core of data handling and storage, a short introduction into database concepts and structure is given. A description of various inputs to databases, including an important explanation of metadata and its value, is presented, as well as possible outputs from such databases. The concept of an 'integrated database environment' is briefly explained, and guidelines for future users and creators of data handling and storage systems are presented.

9.2 HISTORICAL OVERVIEW

Data have been collected throughout history. Prior to the advent of computing technologies, data were physically stored on material such as stone (and later paper), providing relatively permanent storage. Drawbacks of such an approach were numerous: storage was very expensive; only a limited number of users could have access to those data; little information about possible accuracy of the data was available; later processing of the data was difficult. However, there was one major advantage: stored data could be used even after a few millennia.

Computerization has fundamentally changed data handling and storage. The first computer applications were just copies of the old 'paper' procedures: data were mostly entered by hand and computers were used just as recording devices. Data and their formats were strictly function-specific and hence the number of users and applications of the data were limited. Longevity of the hardware and data formats was initially very short (usually less than a decade).

During the last few years, development in the field of computerization and communications has lead to a change in the paradigm of data use. Large databases are now linked with automatic monitoring equipment, computers are connected into networks and grids, diverse users and applications can share data and true integration of different water related systems is possible if not commonly practised. However, although the technology is largely mature, there are still some issues to be resolved, mostly regarding data lifespan and security.

9.3 DATA FLOW AND DATABASES

Guidance provided in this chapter is ordered to closely follow the typical data flow in water related data systems. Figure 9.1 graphically presents the data-flow process, from a source of the data (see Chapter 7) where monitoring equipment is used to measure certain quantity and to add some basic description about measuring conditions (metadata or GLP, **g**ood **l**aboratory **p**ractice), through Chapter 8 where measured data are validated and a measure of data quality or accuracy is added to the metadata (shown in Figure 9.1 as QA, **q**uality **a**ssurance). This chapter discusses data storage and manipulation keeping in mind the final users and the data applications (discussed in Chapter 10).

As can be seen from Figure 9.1, the **d**atabase **m**anagement **s**ystem (DBMS) is at the core of contemporary data handling and storage systems. The DBMS is a suite of elements consisting of the database and accompanying programmes for data input, output, control, data recovery, and so on. There are a number of commercially available DBMSs, made for different computer platforms. Regardless of the database structure, all DBMSs have the following characteristics (Prodanović, 1997):

- *Data independence*: User programmes that communicate with the database through the DBMS are independent of access strategy, physical structure and characteristics of the physical storage device. The DBMS acts like insulation between the programmes and database.

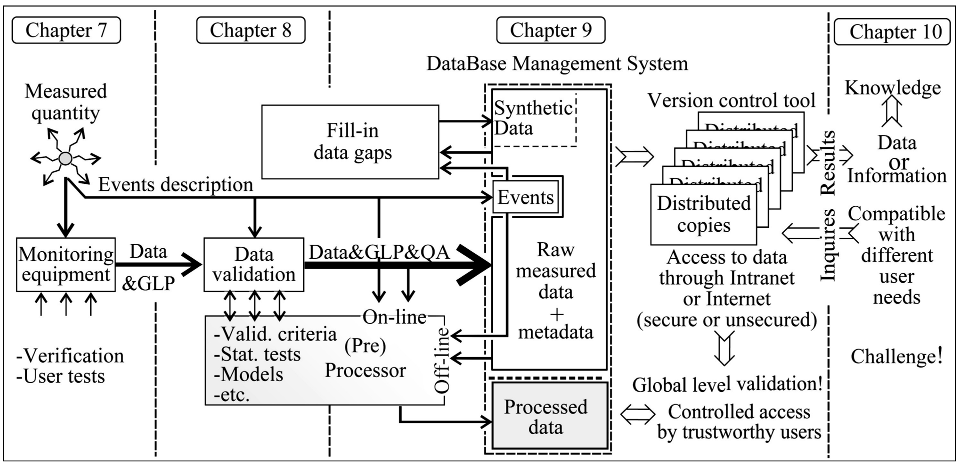

Figure 9.1 Various stages of the data flow process that leads to knowledge creation (relevant chapters in this book noted above; GLP = good laboratory practice, QA = quality assurance)

- *Data structuring*: Use of a large database would be a nightmare if the data were not structured in a meaningful way. The DBMS should take care of data structure and should be able to express the most complex relationships which may exist between the data.
- *Storage, validation and recovery*: The database is a 'shareable' resource and thus potentially vulnerable to errors (or event deliberate attack). The DBMS must validate the data before storing it and should be able to perform full recovery of the database. A mechanism for transaction tracking allows the DBMS to maintain the integrity of the database. Permanent storage of the data is the DBMS's primary role, so backup procedures and hardware redundancy are compulsory.
- *Redundancy control*: There are two types of redundancy. The first one is within acquired data when the same quantity is measured using two or more sensors, in order to assess the uncertainty during data validation phase (see Chapter 8, validation Criteria 6 and 7). The second is redundancy within the database itself, when several copies of the same data exist in different tables in order to improve database performance (speed and/or security). The DBMS can control only the second type of redundancy, by keeping all copies up-to-date.
- *User views and security*: In general, a DBMS is designed to be used by a number of users who may access the same database, and each user can work with certain parts of the data. However, when the database is used to store the raw measured data, it is not a good practice to allow access for all users. Although the users could be divided into different groups by the DBMS, and each users' group or individual user can have a number of security restrictions (username, password, ownership, right to read data, right to change data, etc.), it is much better practice to limit access to the main database to a few authorized users, and to have separate databases for all other data users. The DBMS will take care of making the distributed copies and keep the copies up-to-date. If the user discovers that some data within the main database has to be changed, he/she will need to send a request to an authorized user to complete the change.

The common structures used for databases are: hierarchical, network, relational and object-oriented. The first two structures were more popular during the 1970s and 1980s, while the relational database structure is dominant these days. A full object-oriented structure is possible but needs a lot of computing power, so an object-oriented approach applied over the relational database is more popular, mostly because it adds the 'object' flavour to the well established relational model.

Each of the four database structures has some benefits and drawbacks, depending on the type of data to be stored. Hierarchical and network structures are the best when clear relation between stored elements exists (in some textbooks those structures are called 'navigational' structures). Those structures allow high-speed access to large volumes of the data, as well as easy update and inclusion of the new data.

The relational database structure is the most popular for general data storage. It uses the concept of linked two-dimensional tables. To store or to extract the data, queries are used; in most relational databases, the standard query language (SQL) is used. The main disadvantage of the relational structure is the time-consuming search along the tables, although indexing and creation of key-elements are used to accelerate the process. On the other hand, due to its flexibility, the relational databases are the best choice for general data storage systems.

The object-oriented (OO) approach is a contemporary trend in information technology. These systems are more modular, easier to maintain and easier to describe than the traditional, procedural systems. However, object orientation is often used without clear definition of what it actually provokes. One of the definitions of object orientation is that an entity of whatever complexity and structure can be represented exactly as one object, in contrast to other database structures where an entity has to be broken down to the lowest level. An OO database system will store and manipulate objects as they are, that is, with their state, behaviour and identity maintained, and not as a collection of their components. However, actual database implementation is not exactly as the definition might imply: each object is really stored as a number of basic entities, together with the topological relationship and methods used to manipulate the object; all the complexity is hidden from the user, in essence encapsulated. In general, the resulting OO database will be much bigger than a relational database, with a consequent reduction in speed.

9.4 INPUTS TO AND OUTPUTS FROM THE MAIN DATABASE

The main database has to handle numerous data inputs and output requests. The data inputs in most cases are in the form of data streams, coming from different communication channels using different hardware. To handle this, some kind of standardization is needed for data format and communication protocols.

Basically, there are five types of inputs into the main database:

Type 1: True raw data as the result of some measurement. Such data could be:

- *the value of a certain quantity that is fixed in space (e.g. a point value, values along a line, or values in space within the fixed frame) and variable in time*. Inputs are mostly in the form of time series with constant or variable time steps, or sporadic measurements performed at certain instants in time. Examples are numerous, for example, flow rate in a pipe, air temperature at a meteorological station, water level in a borehole. Such data in most cases could be automatically imported, either through a direct link (e.g. GSM, radio link, optical link) or by manual file exchange.
- *the value of a certain quantity which may change in time and in space in a way that no fixed space frame can be used*. The volume of such data can be extremely large, depending on the spatial and temporal resolution (see Chapter 5). Some biological measurements fall into this category, such as algae blooms and fish behaviour and migration patterns.
- *positional data and attributes such as those about infrastructure systems, objects on the ground, cadastral data, soil characteristics, positions of known leaks from water pipes, borders of retention ponds*. Most of these data are best handled by a GIS (geographic information system) as part of database management system. However, it is important to keep in mind that existence of the features, their position and/or attribute values are not constant in time. For example, if a hydraulic model of a water network is made based on GIS data supplied by a water utility, and continuous monitoring of pressures and flow in the system is used to calibrate the model, it can happen that the network layout is not continuously up to date. During the calibration period, the maintenance crew from water utility may close or change certain pipes in network.

Type II: Metadata. Metadata is literally data about true data. It is an essential part of database, since metadata is critical for integration between different databases and users.

Metadata is a term that is understood in different ways by diverse professional communities that design, create, describe, preserve, and use information systems and resources (Gilliland-Swetland, 2004). Until the mid-1990s, 'metadata' was a term most prevalently used by communities involved with management and interoperability of geospatial data, where it referred to a suite of industry or disciplinary standards as well as additional internal and external documentation and other data necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system. The technical committee on Geographic Information/Geomatics (TC 211) has established the standard ISO 19115 'An International Metadata Standard for Geographic Information' that can be found at http://www.iso.org/ (Accessed 02 July 2007.).

More broadly, metadata is *the sum total of what one can say about any information object at any level of aggregation*. The information object is anything that can be addressed and manipulated by a human or a system as a discrete entity; it can be comprised of a single item, or it may be an aggregate of many items. In general all information objects, regardless of the physical or intellectual form they take, have three features: *content*, relating to what the object contains or is about, and intrinsic to the object; *context*, indicating the who, what, why, where, how aspects associated with the object's creation and extrinsic to the object, and *structure* relating to the formal set of associations within or among individual information objects and either intrinsic or extrinsic to the object.

Since the metadata become important in development of networked digital information systems, a broad concept of the metadata has been broken down into distinct categories that reflect key aspects of metadata functionality: administrative, descriptive, preservation, use, and technical metadata. However, in most large databases [or data repositories, such as San Diego Supercomputing Center, www.sdsc.edu (Accessed 03 July 2007.)] the metadata must still be restricted to a specific format, such as the Dublin Core [http://dublincore.org (Accessed 02 July 2007.)], in order to make it usable.

Which categories of metadata are important for integrated urban water systems? The *administrative part* of the metadata takes care of acquisition information (and data accuracy), use rights and tracking of copies made, location information, version control, audit trails created by record-keeping system, among other tasks. The *descriptive part* describes or identifies information resources: cataloguing or indexing records, hyperlinked relations, annotations by users. The *preservation part* tracks actions taken to preserve physical and digital versions of resources (data refreshing and migration to new standards). The parts of the metadata that cope with *data usage* have to track the users and use of data, to select appropriate preview types. Finally, the *technical part* keeps a record of data formats, compression ratios, scaling routines, security issues (encryption keys), availability and response time for remote data sources, and so on.

When creating the database management system with metadata support, it is important to note that all input data, regardless of source, have a certain level of the uncertainty that is not constant over time. For example, daily readings from a turbidity sensor will have increasing uncertainty with the time elapsed since the last calibration (see Chapter 8). Similarly, older spatial data will have higher uncertainty, due to less accurate measurement equipment (Burrough, 1996).

As presented in Figure 9.1, the monitoring equipment and a separate data validation process should automatically create and add the metadata to the measured values. The procedures and metadata descriptions will vary with the type of measured variable. The minimum set of the metadata would include: type of monitoring equipment used, time and date of measurements (single measurement, or start of data collection), sampling conditions (sampling rate, start/stop conditions, alarm levels, etc.), sensor used, calibration data, environmental conditions (temperature, humidity, etc.), sampling place (from GPS data or manually entered), accuracy assessment, results of statistical tests (standard deviation, peak values, mean value trend, etc.) and results of other, well documented validation criteria. Since the metadata should help in data sharing among different users, it is essential to use well known standards wherever applicable (SEDAC, 2005). For example, the World Meteorological Organization's ISO 19115 guidelines [http://www.wmo.int/web/www/WDM/Metadata/documents.html (Accessed 02 July 2007.)] is a useful reference (GRDC, 2005). The use of specialized hydrological 'markup languages' can also facilitate easy integration of large hydrological databases, providing effective data sharing over the internet (Piasecki and Bermudez, 2003).

Type III: Event descriptions. These are generally separated from the regular flow of measured values. Strictly speaking, event descriptions could be considered as a part of the metadata and will be handled with the same metadata-handling routines. However, in most cases, event descriptions will be manually generated messages with a more loosely defined structure, such as photographs, sketches, sound records, or other data collections that are used to better describe the data and conditions during measurements. Under event descriptions one can place the explanation of tests conducted on a piece of system monitoring equipment, such as the exact date, time, and duration of a test and the volume of water poured into the siphon of a rain gauge, (data from which the calibration can be checked, and the volume of water added can be removed from the recorded time series of rainfall intensities), the description of measuring conditions that influence the calibration data (for example, bed silting in the critical depth flume), the changes in the surrounding environment (e.g. increased number of directly connected roofs to the sewer system within a catchment) or other important explanations necessary for later use of the measured data.

Type IV: Synthetic data. During the process of data acquisition, gaps in the recorded series can occur due to problems in communication lines, malfunction of monitoring equipment or other reasons. Depending on measured quantity and possible data redundancy, it is possible to calculate the missing data and fill-in the gaps (see Chapter 8). For example, if three quantities are monitored, i.e. water level in a reservoir, the flow and the pressure in the outflow pipe near the reservoir, a very strong hydraulic relationship can be established among those quantities. Therefore, if the pressure logger runs out of battery charge, the missing part of the data can be easily calculated and stored within the database. Of course, the calculated (rather than directly measured) section of the time series should be marked, noting within the metadata that these are synthetic data. The calculation can be performed off-line either during the validation phase or later during database use (see Figure 9.1).

Type V: Processed data. Apart from the raw measured data, the input into the main Database can be both pre-processed and processed data. Processing can be done online, during raw data capture and online validation, or off-line (most commonly). *The imperative is to separate the raw data from the processed data.*

The pre-processed data can be described as another view on the same dataset, without new information in it. In most cases, pre-processing will consist of data conversion from the raw measurement units to scientific units using calibration data, removing trends, removing high frequency noise, or similar. Pre-processing can also include some kind of data compression, without any loss of data (with simple Run-Length-Encoding schemes, or more complicated schemes with sliding window-like LZ77 derivatives used in ZIP), or with loss of higher frequency components (like JPG algorithm for storage of images). Although the pre-processed data holds the same information as the raw data, it is critically important that the user ***never remove*** the raw data from the database. Safeguarding the raw data will allow the user to return and correct any errors observed at any later stage of data use.

The processed data will create new information out of the captured data. For example, using the frequency domain analysis of the pressure fluctuations and transients in pressurized pipe, the position and extent of leakage could be assessed. Or, total leakage from one district could be obtained by analysis of the difference between mean day flow and night flow. However, the role of the main database is the storage of the raw data, and it should not be overloaded with the processed data. The primary processing work should be done on distributed copies of the selected dataset. This reduces both the workload on the main database, and the risk of damage to the data.

Only a limited number of users should have access to the output side of the main database (in Figure 9.1 those users are called 'trustworthy users'). It is important to separate the main database, which holds the measured data and metadata, from numerous users with their unpredictable requests. The approach that should be applied is to have trustworthy users perform development and maintenance of the database, and have the output copies of selected datasets (or the whole database) accessible by other users. By limiting the public access to the main database, the security issues are much easier controlled.

Links from the main database to copies prepared for general usage should be *unidirectional*. If something is to be changed in the main database, it must first be verified by authorized users. These users must also ensure synchronization of the data between the main database and any derived copies. *Version control is important, since the users will work with different models, assumptions, or boundary conditions, and will produce different outputs using the same input data*. Also, version control has to cope with the possibility that working copies of the data from the main database could be different, since different selection criteria can be applied, or different data compaction algorithms used. For example, one copy of a rainfall dataset may have only daily mean values, while another may have only events with rainfall intensity higher than a certain threshold (see Chapter 8).

9.5 INTEGRATION OF DATABASES

The main database with its management system (as described and presented in Figure 9.1) is just one system for data handling and storage within the complex field of IUWM. In a complex water related organization several DBMSs will exist and will be linked.

The water utility system in Figure 9.2 provides one example of integration. A number of smaller DBMSs handle data and information related to water: LIMS is the Laboratory

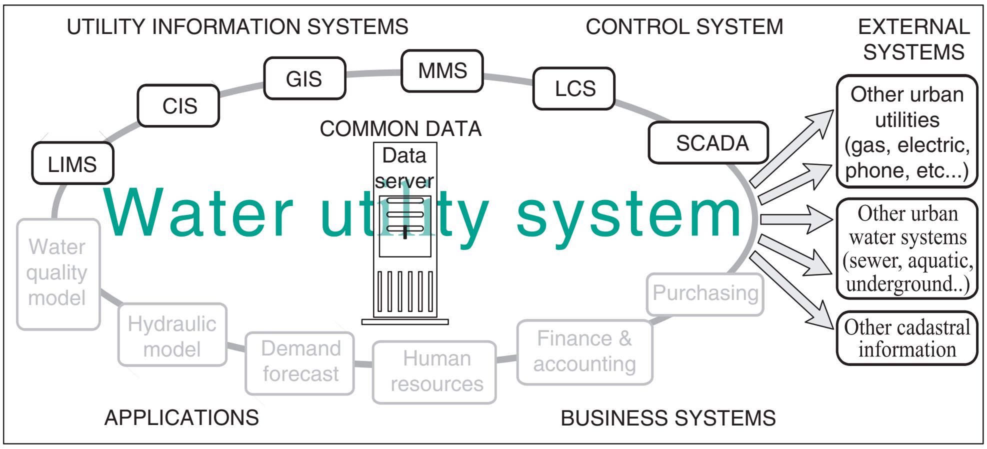

Figure 9.2 Example of data/information flow within the complex system of integrated urban water management (LIMS = Laboratory Information Management System, CIS = Customer Information System, GIS = Geographic Information System, MMS = Maintenance Management System, LCS = Leakage Control System and SCADA = Supervision, Control, data Acquisition and Data Analysis)

Information and Management System; CIS is the Customer Information System; GIS is the Geographic Information System; MMS is the Maintenance Management System; LCS is the Leakage Control System; and SCADA provides online network Supervision, Control, Data Acquisition and Data Analysis. All these systems are of vital importance for real-time functioning of the water utility, both technically and financially. Each system will collect data and will share subsets of those data with other systems. There should be, however, one common data repository system acting as a central supervisory unit to handle storage of the metadata (links and explanations of the type of the data and their available formats in the given DBMS), serve as a communication centre for links to other external systems, and handle the privileges and rights for access to specific data.

To achieve full integration the information system of water utility should be linked with external systems, such as urban utility systems with spatial (positional) data regarding gas, electrical and telephone lines and cadastral systems where legal information about consumers and design consumptions (ie. the level of consumption on which the system design is based) can be found, and other water-related systems such as the sewer utility, underground water and surface water. Communication among those systems should all be passed through the central supervisory unit, which will process the requests, check access rights, consult its metadata database for the actual location of the searched data and ask the local DBMS to make a copy of the required data subset.

9.6 CONSIDERATIONS FOR DATABASE USAGE WITHIN AN IUWM FRAMEWORK

There is a considerable worldwide effort to integrate fundamental information and data in the field of water management. Numerous initiatives and projects have been conceived after recognition that data integration is urgently required if integrated urban water management is to be realized (Maurer, 2003, 2004).

Since data sharing among urban water components is imperative for the database to function in a sustainable long-term manner, its development has to be thoroughly analysed and investigated. The available standards for databases and metadata management (SEDAC, 2005) should be followed closely to facilitate data-sharing. In addition to the international standards, some considerations and recommendations that should be taken into account during the process of database creation can be summarized as follows:

- *Store the input data in raw format as received from monitoring equipment*. Store separately the used SI conversion factors and calibration data. Store the raw calibration data for each sensor and keep a record of previous calibrations to be able to analyse long-term stability of the sensor.
- *Collect adequate metadata*. Data in any database are useless without accompanying metadata (geographic position, sensor identification, responsible person, link to calibration data, assessed uncertainty, etc.). If possible, all potential future users of data should define their prerequisites for minimum and optimum metadata (Vyazilov et al., 2003).
- *Input data should be separated from information produced by (pre-) processing* – ***never mix*** the two (data/information version control). The raw data are unique, while the information developed from data processing can have different versions, depending on rules used in (pre-) processing. Metadata record the details of the various versions.
- *Redundancy in the input data is important, but it shouldn't be misused*. Redundant data will help in assessing data accuracy, in filling the gaps of missing data, and in crosschecking.
- *Hardware redundancy of the main database is compulsory*. A number of techniques exist, including disk mirroring, processor mirroring, and dual power supplies.
- *In the (pre)-processing stage, all measured time series should be placed on a common time line*. For example, if flow and pressure are monitored, and pressure readings were taken at 08.23.15, 08.28.29, 08.32.07, … (i.e. with 5-minute sampling rate) and flow readings at 08.22.11, 08.32.09, 08.42.18, … (i.e. with a 10-minute sampling rate), interpolate both readings to the same time interval, for example, 08.20, 08.30, 08.40, … and so on. The interpolation is a part of preprocessing, so again, the raw data should be retained at all times. Note, however, that the results of such interpolation algorithms should be checked for data degradation, especially in the cases where gradients of the measured quantity are steep.
- *Data compression should be used, but with great care*. If the consequence of such data compression is a loss of information, the input raw data should not be compacted since data reduction will reduce overall accuracy. On the other hand, lossless compaction of the input raw data can be done, but it will reduce the reliability. If an error in the compacted data occurs, in most cases the whole time series will be lost. On the other hand, if an error occurs in a simple input raw data file, most of the data can usually be saved.
- *To achieve longevity of the main database, great care should be devoted to the selection of the software platform*. Often changes in software versions can lead to expensive database maintenance. Consider usage of open source platforms, since higher initial effort in database creation will generally produce lower running costs.

- *Use internationally established standards wherever possible*, for example, for storage of raw data (AGIT, 2003), for metadata (DCMI, 1999), for formating exported databases (GISHydro, 2005), for internal and external communication, for navigation within the database (SQL language) and so on.
- *Do not digitally pollute the database by storing unnecessary data*. The fact that current technology allows recording of high-resolution pictures and movies or time series acquisition with extremely fast sampling rates, does not mean that the database should be loaded with such data. If those data are redundant, contributing little to the primary purpose of the database, they will simply degrade its performance.
- *Portability of data format has to be maintained from the data source (monitoring equipment) to the data storage*. To achieve portability of the format means that the needs of all potential future users and models should be well considered from the outset. Later conversion of the data in the most cases is not straightforward.
- *Consider thoroughly the implications of temporal and spatial scales on data storage requirements* (see also Chapter 5). Data reduction techniques can be used, but this will increase the overall error, so careful selection of thresholds in data compaction should be done.
- *Database management system must address security issues*. Data security has two aspects:
  - data encryption, as secure data storage and control of access to data, and
  - prohibition on changing measured data.

  Only small group of authorized users should have full access to all database resources. The main database should not be widely visible (through the internet, for example), so the main computer should not have an internet protocol (IP) address. One-way communication from the main computer to publicly accessible copies of the database should be used.
- *Appropriate cross-checks should be put in place to deal with possible discrepancies in data stored at different locations*. If intelligent instrumentation for data acquisition is used (see Chapter 7), the raw measured data will be stored at two locations, namely, in temporary memory within monitoring equipment and in permanent storage within the main database.
- *Performance indicators for data-storage systems should be established and continuously monitored*. Indicators could be, for instance, on the rate of database enlargement, number of megabytes per asset, number of database users per month, and so on. Such indicators can be used for maintenance and growth predictions.
- *Database maintenance and cleanup should be done on a regular basis*. There are two important issues that should be considered:
  - Which data should be kept and for how long, and which data should be thrown away? Clearly, within the database, some raw data sets should be kept 'as is' permanently, and other raw data sets are needed only for a limited period of time (they are disposable after completion of certain projects) and then they will be compacted, or kept only as statistical values (averages, extremes, etc.), or completely deleted. The procedure for regular database inspection has to be established in which every year (or every second year) the performance indicators on data usage should be calculated for data sets marked as 'disposable' and appropriate action taken.
  - Long term data storage or the persistence of used data storage. Here, three points are to be resolved.

- How long will the storage media hold the information? What conditions and what kind of maintenance is needed to preserve the recorded data on media (humidity, temperature, rewinding or refreshing of tapes)? In general, magnetic media can be used for a few decades, optical media for 50 years to 100 years. If data are needed for centuries, then appropriate migration plans are needed.
- Changes in data storage standards and maintenance compatibility with previous standards? The time scale here is much shorter than the problem of longevity of the storage media. Again, an appropriate migration plan should be developed.
- Database management software is constantly changing and improving. To keep pace with this, use software solutions which are as standard as possible, preferably with open-source components where possible and well supported.

REFERENCES

AGIT. 2003. *Guidelines for the Archiving of Electronic Raw Data in a GLP Environment*. Working Group on IT. www.glp.admin.ch/ (Accessed 02 July 2007.) (GLP-ArchElectRaw Data1_0.pdf).

Burrough, P.A. 1996. Natural objects with indeterminate boundaries. P.A. Burrough and A.U. Frank (eds), *Geographic Objects with Indeterminate Boundaries*. Bristol, UK, Taylor and Francis.

DCMI. 1999. *Dublin Core Metadata Element Set*, Version 1.1, Reference Description. Dublin Core Metadata Initiative. dublincore.org/documents/1999/07/02/dces/ (Accessed 02 July 2007.).

Gilliland-Swetland, A.J. 2004. *Introduction to metadata – setting the stage*. Los Angeles, USA, J.P. Getty Trust. http://www.slis.kent.edu/~mzeng/metadata/Gilland.pdf (Accessed 03 July 2007.).

GISHydro. 2005. *Hydrologic modeling using GIS*. Austin, TX, US, Center for Research in Water Resources, University of Texas at Austin. www.ce.utexas.edu/prof/maidment/gishydro/home.html (Accessed 02 July 2007.).

GRDC. 2005. Metadata. Global Runoff Data Centre. http://grdc.bafg.de/servlet/is/2377/ (Accessed 02 July 2007.).

Maurer, T. 2003. Intergovernmental arrangements and problems of data sharing. Paper presented at Monitoring Tailor-made IV Conference on Information to support Sustainable Water Management: From Local to Global Levels. St. Michielsgestel, The Netherlands. grdc.bafg.de/servlet/is/3997/ (Accessed 02 July 2007.).

———. 2004. Transboundary and transdisciplinary environmental data and information integration – an essential prerequisite to sustainably manage the Earth System. Paper presented at Online Conference on INDUSTRY IDS-Water Europe 2004 www.idswater.com (Accessed 02 July 2007.).

Piasecki, M. and Bermudez, L. 2003. Hydroml: conceptual development of a hydrologic markup language. Paper presented at the XXX IAHR Congress, 24–29 August 2003, Thessaloniki, Greece.

Prodanović, D. 1997. Introduction to Geographical Databases, Development and Maintenance of Database for Urban Infrastructures, Database Matching with Simulation Models, and Data Structures for Physically Based Models. E. Cabrera and Č. Maksimović (eds), *Sistemas de informacion geografica (GIS) aplicados a redes hidraulicas*. Valencia, Spain, Grupo Mecanica de Fluidos, Universidad Politecnica de Valencia.

SEDAC. 2005. Metadata. Socioeconomic Data and Application Center. sedac.ciesin.org/metadata (Accessed 02 July 2007.).

Vyazilov E., Mikhailov, N., Ibragimova, V. and Puzova, N. 2003. Metadata as tools for integration of environmental data and information production. N.B. Harmancioglu, S.D., Ozkul, O. Fistikoglu and P.G. Fistikoglu (eds), *Integrated Technologies For Environmental Monitoring and Information Production*.