Chapter 10

Use of data to create information and knowledge

D. Prodanović[1]

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

10.1 INTRODUCTION

Collected data constitute a basic source of information. They form a window that we use to look at our environment (Singh et al., 2003). Data are used for a large number of activities (see Chapters 1 and 3), such as for assessment of a system's performance, calibration and verification of Integrated Urban Water Management (IUWM) models, real-time control of IUWM systems. The portability, presence of metadata, and their reliability are critical to the process of data translation into required knowledge. The steps and procedures used for turning data into information, and then finally into knowledge are explained in this chapter.

10.2 DEFINITION OF TERMS

Data are classically defined as the basic building blocks of human knowledge and consist of separate, uncorrelated raw facts (IBM DB2, 2003). *Information* is data endowed with relevance and purpose. Using relationships among the original facts (raw data) a meaningful context is given to data. *Knowledge* is a step further from information. Knowledge is created only when human minds incorporate (accept) and act on information through *decisions*. Information can be created from the data using different, mostly computerized techniques. In the process of knowledge creation, however, technology can only help humans to select appropriate information, but *human beings must convert the information into knowledge*.

The classic definition of data implies that the raw datum by itself delivers no benefits to the final user. Knowledge is needed to decide on certain action (Figure 10.1). A massive dataset may even be a distraction if not processed into information. Also, information does not lead to the decision-making process until people learn it and accept it.

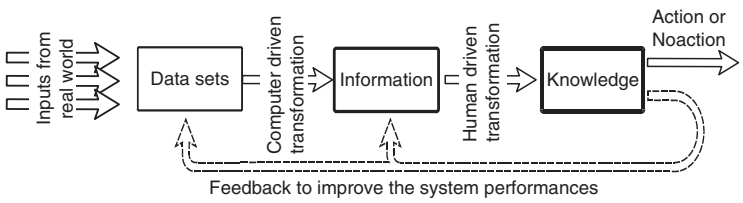

Figure 10.1 Data must be converted into knowledge in order to be useful

According to some authors (Santos and Rodrigues, 2003; Thearling, 2007) the knowledge discovery process presented in Figure 10.1 consists of a seven-step sequence (Han and Kamber, 2001; Babovic et al., 2002):

(1) *Data cleaning* (as explained in Chapter 8), to remove noise from data and inconsistent data sets;
(2) *Data integration*, to combine different sources of data;
(3) *Data selection*, to retrieve relevant data for analysis: an appropriate data sampling strategy has to be defined;
(4) *Data transformation*, to process data into a form suitable for data mining, through dimensional reduction using aggregation operations;
(5) *Data mining*, to identify patterns (relationships, events or trends, which may reveal both regularities and exceptions among data) and enable model selection;
(6) *Pattern evaluation and interpretation*, to identify interesting patterns representing the knowledge;
(7) *Knowledge representation and usage*, to represent the gathered knowledge to the user and its use in decision making.

All these steps assume that the data about the real world are already acquired and stored within available databases, actually the most costly and technically complex task as it involves conversion of data from our 'analogue' world into a digital representation. Some authors call those preparatory steps 'data consolidation' (Thearling, 2007).

Figure 10.2 extends the process of data transformation into usable knowledge given in Figure 10.1. It shows the feedback process used to improve the overall system

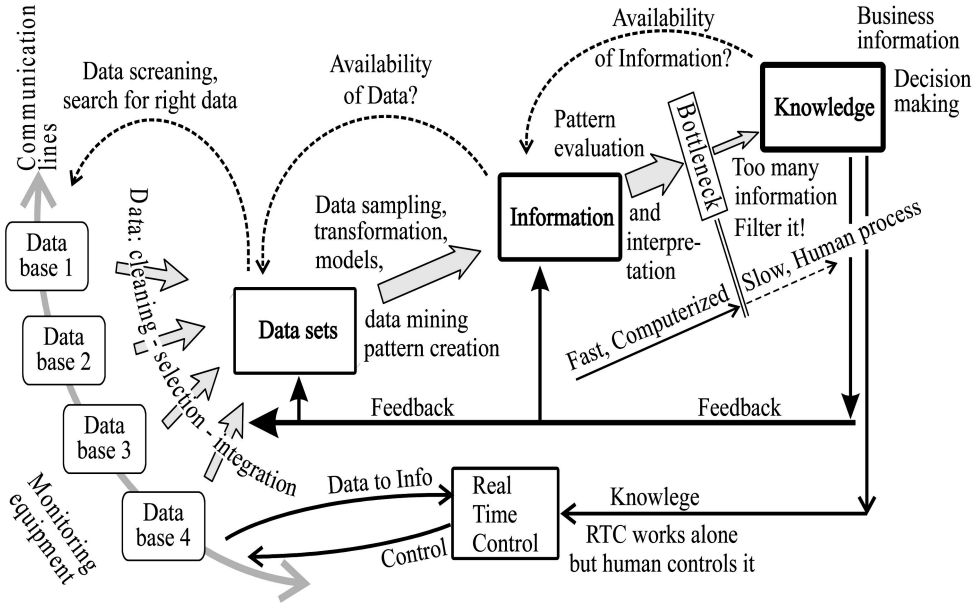

Figure 10.2 Conversion of collected data to knowledge as an interative process

performance at each step. Two important considerations are underlined on the upper left part of Figure 10.2:

- *Data screening*: How can we know *what data are available* and *where they are*? In complex systems, like urban water, a number of organizations are in the business of data collection. Existence, accessibility and availability of such data become an important 'source of information' in itself and are key components of making and keeping those data useful.
- *Searching for the right data*: How can we know that the *right data* were collected and stored from the *right place* and at the *right point in time*? For knowledge creation, *no information* is much better than *bad information*. The single most elusive problem associated with transformation of data to knowledge is identifying errors while gathering accurate information. That is why data validation (Chapter 8) is so important and why metadata associated with accuracy assessment must be stored together with raw data.

The steps from data capture and data cleaning through pattern evaluation are nowadays carried out using computers. Large quantities of data can be easily processed and vast amounts of information created. However, the path from information to knowledge still has one important 'bottleneck', human beings. Humans have to be able to accept all those patterns and information in order to compile it into knowledge. This bottleneck we must be aware of when using automatic filtering of irrelevant patterns and identification of interesting information. Often the capacity of computers to produce patterns and information exceeds the capacity of human beings to interpret them.

The conversion of data to knowledge, as presented in Figure 10.2, implies that the users of data are managers and politicians that will use expertise of scientists in knowledge creation, to inform business and political decisions. But in many water-related organizations, most data are still measured in order to control and manage the system in real time. **R**eal-**t**ime **c**ontroller (RTC) devices designed to act upon a certain state of input parameters and measured data. In this way, RTC systems can be seen as 'data to knowledge converters', with more or less artificial intelligence (Figure 10.2), however, RTC systems are still controlled and programmed by humans.

10.3 FROM DATA TO INFORMATION

While data is simply the product simply of collection, information is defined by the content of the data, which is meaningful to the user and relevant to the question of problem being addressed. Data exist from the moment it is captured and stored, while information exists only if it is useful to some audience or decision makers in the most general and inclusive sense.

The database management system therefore has to be designed to support information retrieval. Portability of stored data and the presence of metadata are prerequisite, as described in Chapter 9. As different urban water systems are integrated, portability of the data becomes the biggest challenge, since different data users with different needs will have an access to local databases.

A scheme for information retrieval in an integrated environment is presented in Figure 10.3. The data user (presented as 'Workstation') will ask the 'common data

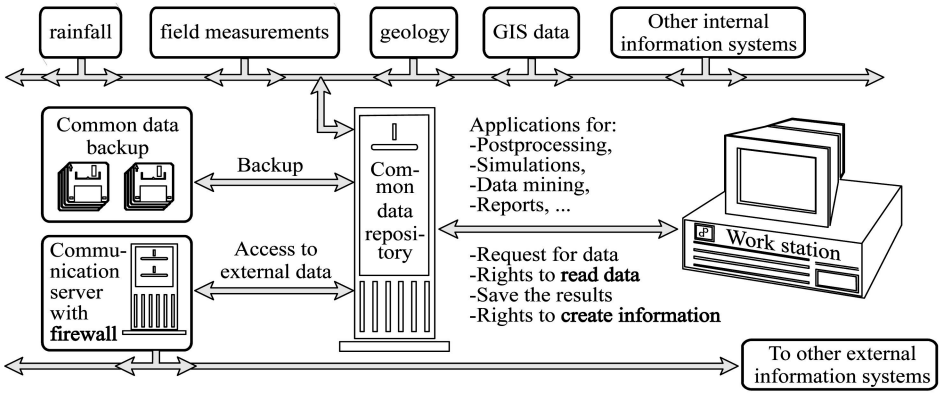

Figure 10.3 Schematic of an integrated database and data-processing system

repository' (the database with metadata about the raw data and storage locations) what data are available and where they are. Automated searching and sorting tools have been developed to keep track of where data are stored (e.g. SDSC, 2004). While searching for data to be analysed, care should be taken on systems with **d**ata-**r**ich **i**nformation-**p**oor **s**yndrome (DRIPS) (Maksimović, 1999).

An authorized user will extract data and then perform different techniques to develop information from this data. This may include running simulation models and performing calibration of them, as well as, say, simply visualizing the data. Any derived datasets should be accompanied by metadata that describes exactly how it was derived.

Any information extracted from the data should have the following properties (Fedra, 2003; Harmancioglu, 2003):

- It should be *timely in relation to the dynamics of the problem to be addressed*. For example, when a pipe bursts, information on what valves to close in order to cut-off the flow is needed urgently. On the other hand, information on population growth, useful to predict future water needs is required over longer time frames.
- It should be *available when it is needed*. For example, predictions need to be made in time to react to an event.
- It should *maintain the accuracy of the data used to create it*. Proper creation and maintenance of metadata is necessary to ensure that only accurate data are used.
- It should be *accurate and precise in the frame of the information requirement*. For water supply, for example, certain precision will typically be a legal requirement, if it to be used for billing. On the other hand, excess precision will unnecessarily consume storage.
- It should be *easy to understand*. The information is just a step toward its acceptance by humans and subsequent conversion into knowledge. Understanding is prerequisite to this process.
- It should have a *format expected by and adequate* for the audience and users.
- Its *context should allow and facilitate interpretation*. Information should not be ambiguous; it must have a unique meaning.

- It should be *easily accessible* (free or not), that is, *expensive in relation to the implied costs of the analysed problem*. For example, if a water utility manager needs data on total consumed electrical energy in the previous year for rough assessment of power needed for the current year, two sources of data could be used: the first one, much faster and cheaper, will be by summing all consumers' electricity bills. The second one could be by taking readings from all **s**upervision, **c**ontrol, data **a**cquisition and **d**ata **a**nalysis (SCADA) systems, sorting the consumption according to different types of electrical devices and types of consumers, checking of electrical tariffs. The second approach is much more elaborate, will give better insight, but is also much more time cleaning and more expensive.

The user has a number of information technologies available to process data (Shaimardanov et al., 2003). These technologies, none mutually exclusive and often combined, include:

- GIS, a set of tools for data retrieval, selection, manipulation and preview;
- statistical analysis for re-processing of data, data aggregation and subsampling;
- data mining for the automated search for certain patterns within data or for certain theoretical relations;
- simulation models used alone or within data mining;
- the internet for data acquisition, searching for data and knowledge, the dissemination of obtained information, and the distribution of the data mining work load to grid computers, and
- *Object orientation* as an encapsulation of the above methods.

10.3.1 Geographic Information Systems

Geographic **I**nformation **S**ystems (GIS) facilitate the capture, storage, retrieval, analysis and display of geographic and spatial data (Burrough, 1993). The basic working forms are maps, that is, all data within GIS are spatially (geo-) referenced. GIS works generally with very large volumes of data and uses complex concepts that describe geometry of objects and relationships between them. The spatial objects have attributes or properties that could be a function of time but are independent from the location in space. The space in GIS is defined through layers and the relation between objects, so user can easily locate them. A GIS has two major roles: one is data capture and storage (see Chapter 9), and the other is data integration, analysis, integration with external models, as well as presentation and dissemination.

The GIS handles inquiries from various groups of users with different views of the same spatial data and with different processing needs (Prodanović, 1997). The key difference between GIS and other automatic cartography systems (AM/FM systems) or CAD programmes is their ability to integrate geo-referenced data from a number of layers (and different sources) and to create new information out of existing data. Successful GIS usage depends on:

- data availability (at the appropriate scale for the problem);
- adequate concept of internal data organization;

- existence of metadata to allow integration of the available data;
- a decision model for users that will integrate the gathered data, subsample them according to some criteria, transform them using the selected model, and create new information or patterns,
- criteria for model evaluation, where the GIS visualization functions will help in presentation of model outputs (in time and in space), thus increasing the speed of information to knowledge transfer.

According to some authors, GISs were originally developed as an operational tool: to manage vast amounts of spatial data. Now GISs are shifting from being simple operational tools to being strategic decision support systems, incorporating more powerful analytical techniques (Sholten and LoCascio, 1997) as a result of a number of developments:

- visualization that has evolved considerably: three-dimensional virtual reality functions and multimedia offering greatly improved information evaluation by users;
- communication possibilities that allow easy sharing of data with closed proprietary systems now transformed into widely accepted open GIS (Raper, 1997);
- advanced spatial data analysis methods based on neural networks, genetic algorithms, fuzzy data concepts, interpolation and extrapolation of data;
- hardware progress relative to price, particularly for GPS.

10.3.2 Statistical analysis

The extraction of statistical information is probably the very first thing each data user will do when faced with a new dataset. For example, when concentration of dissolved oxygen (DO) is monitored in a small creek, the user may calculate statistics such as the mean, median, minimum and maximum. Depending on the main database design (Chapter 9), some primary statistical values could even be stored as metadata in the (pre-) processing stage, thus speeding up the possible search by remote users for relevant raw data (e.g. search for records where the mean DO concentration is less than 5 mg/L).

In general, there are three statistical concepts that are used in data analysis (Singh et al., 2003) regardless of the dimensionality of data:

(1) extraction of aggregate characteristics, that is, calculation of mean values, along time or space, specifically the arithmetic mean, median, mode, harmonic mean and geometric mean (with the calculation of mean values reducing dimensionality of the data by one degree);
(2) extraction of variations of individual values from aggregate properties, including the calculation of deviation (mean and standard), variance, coefficient of variation, skewness; and
(3) change of the time/space domain into a frequency/time or wavelet domain. Standard methods of time series analysis include the analysis of frequency distribution of individual values (empirical or theoretical), usage of 'Fourier transform' to extract the dominant frequency components using periodic sine and cosine functions, or the wavelet transform with optimized mother functions.

All three concepts could be used either as data transformation tools to process the data into new information, or as pre-processing tools to reduce dimensionality of the problem.

Based upon the extracted statistical data, it is important to maintain an active feedback with the main database from which the raw data were extracted, as previously discussed. The feedback should provide information to the main database on:

- monitoring equipment, with respect to selected time and spatial resolution (for example, the sampling rate is too slow for the monitored variable, or the sensor range used is too wide), assessed accuracy (more accurate measurements are needed) and availability of metadata from monitoring equipment (requests for more specific metadata);
- new requests for (pre-) processing of the raw data and general work with the metadata; and
- sampling criteria used to extract the data from the main database and to create the remote database accessible by users.

10.3.3 Simulation models

Simulation models are *the representation of physical laws or processes expressed in terms of mathematical symbols and expressions* (i.e. equations). Such models are used as a basis for computer programmes where the effect of changing certain variables on the output result can be examined, for example, the analysis of the effect of daily variation in water consumption on water delivery system.

As the raw data are a representation of the physical world's current state through sampled fragments, the simulation model is used to represent the continuous complex interactions between different variables and processes. The simulation model should respond to inputs as it would in the real world, allowing the user *to interpolate the fragments of measured data into a continuous series*. Such models can also be used for past and future prediction (assuming constancy), and thus become very useful for making decisions about possible management actions.

Simulations models are a powerful and commonly accepted technique for extracting information from available data. The data are used during several stages, including:

- *Model creation*: The concept of the real world simulation is heavily driven by data availability. The old concepts of well-established models are now changing (Maksimović, 1999) towards data-driven physically based models (Drécourt and Madsen, 2001; Katopodes, 2003). With the integration of the data related to all aspects of urban water cycle, it is possible to develop a complex integrated model that will allow full interaction of several systems. A good example is presented within the UNESCO IHP VI Project 3.5.3 'Urban Groundwater Interactions' where three urban water components are linked: water supply, sewer system (channels and trunks), and underground water.
- *Model calibration*: Sufficient amounts of measured user-selected data are needed to calibrate the model. The calibration phase of the simulation model can, in some instances, result in changes in already accepted model concepts.

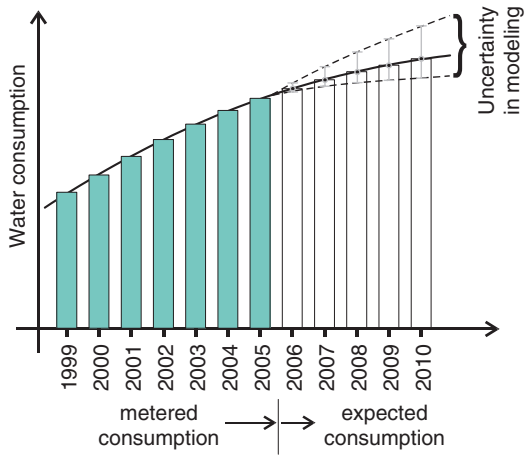

Figure 10.4 Extrapolation of water consumption using a simulation model with a specified level of uncertainty

- *Model verification*: The data used in verification is different from data used during the calibration phase. Model verification can be also done continuously using newly measured data, so changes in the real world that are not implied by the model can be discovered. Through the model verification, a measure of model uncertainty can be established. There are two types of uncertainties that should be addressed (see Chapter 6), specifically, first, how accurately is the true world represented and what are the limits of model? What are the errors that the model will produce using accurate input data, within specified model limits (i.e. when used for interpolations) and outside the limits (if used for extrapolations)? The answers to these questions are directly related to the model concept (whether it is a physical model, conceptual model, simple parametric model, etc.) and data sets used for model calibration. Secondly, the propagation of data uncertainty through the model and interaction with the model's own uncertainty (see Figure 10.4) should be addressed.
- *Model usage*: Once the model is calibrated, the user can 'play' with it, trying different scenarios to learn what might be the response in the real world. An important part of each simulation model is the presentation of the results, converting the information into usable knowledge.

The interactive use of models provides a 'feedback loop', helping to improve understanding of the system. In turn, the user with this new knowledge may determine to collect better data to further improve the model.

Simulation models can also be integrated into bigger, more complex models of the urban water system. An example of several simulation models integrated into the larger conceptualization of the water utility system is presented in Figure 10.5. The demand forecast model, for example, will take samples of data from the Customer Information System (CIS), locations from the GIS, current flow and reservoir levels measurements from the SCADA, and meteorological data from external system. The demand forecast

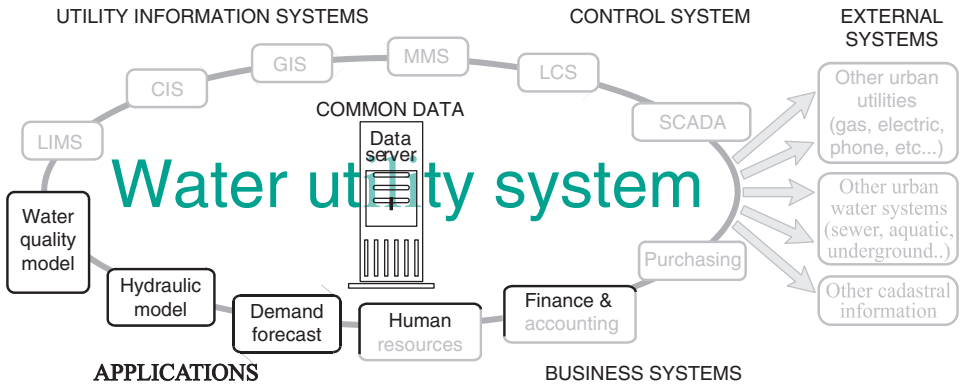

Figure 10.5 Example of applications of an integrated use of simulation models as part of an Integrated Urban Water Management system (LIMS = Laboratory Information Management System, CIS = Customer Information System, GIS = Geographic Information System, MMS = Maintenance Management System, LCS = Leakage Control System and SCADA = Supervision, Control, data Acquisition and Data Analysis).

model is linked with a hydraulic model of water transport through the network, which will use also GIS data and data from the Leakage Control System (LCS). The hydraulic model is closely coupled with the water quality model and calibrated using the results of the Laboratory Information Management System (LIMS). The final goal of all these integrated models is to predict the quality parameters at the customer's connection and to suggest the appropriate actions to maintain those parameters within allowable limits.

A well-established and calibrated model can be used to extract reliable information even from incomplete data sets. If the pressure sensor in SCADA is out of order and its signal is missing, the simulation model can be used to compute a dummy number to fill in the gap within the database, as shown in Figure 9.1. Of course, this entry has to be marked within the metadata as 'simulated'.

Sometimes, the simulation model can be used to reduce the cost of equipment by monitoring a variable indirectly using some other easily measured quantity. For example, turbidity is often used to monitor total suspended solids (TSS) using simple correlation models (Fletcher and Deletić, 2007). If this is the case, it has to be recorded within metadata, since it affects knowledge derived from the TSS readings.

10.3.4 Data mining

The main drawback of using a simulation model for information extraction is that the user must have a good previous knowledge to prepare the model and to prepare the input data for calibration and model usage. When faced with a number of available data series within an integrated urban water database environment, analytical tools need to include intelligent reasoning in computerized data analysis. The general name for all such tools is *'data mining', the automated analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized* (Savic et al., 1999), or, according to Kurt Thearling, the extraction of hidden predictive information from databases (Thearling, 2007).

The term 'data mining' appears in the literature under a multitude of names, which includes knowledge discovery in databases, data or information harvesting, data archaeology, functional dependency analysis, knowledge extraction, and data pattern analysis (Savic et al., 1999). Recent improvements have seen advances in the use of data mining for environmental numerical and non-numerical data, including pattern recognition in spatial data (Wachowicz, 2000).

Data mining can be conducted as:

- *Unsupervized learning* (Roiger and Geatz, 2002) or *unidirected* or *pure data mining* (Savic et al., 1999) or *data-driven mining* (Babovic et al., 2002): A data mining method that is left relatively unconstrained to build models and discover patterns in the data, free of prejudices (hypotheses) from the user. It is thus a true discovery process and is used usually for classification and clustering.
- *Supervised learning* or *directed data mining* or *theory-driven mining*: The user builds a 'learner model' or concept definition based on existing knowledge and understanding of physical processes. Data mining is used to train the model by comparing the known input/output relations. The model is then used to determine the outcome for new input instances.
- *Detection of anomalous data and patterns*: The user applies previous data mining results to analyse anomalous patterns and unusual data elements, that is, those that do not conform to the general patterns found.
- *Hypothesis testing and refinement*: The user presents a hypothesis to the system for evaluation and, if the evidence for it is not strong, seeks to refine it.

Within all the above learning categories, two main data-mining tasks can be identified (Santos and Rodrigues, 2003; Savic et al., 1999), namely,

- *prediction*, a task of deduction using the data to make prediction, incorporating classification, regression and time series analysis; and
- *discovery* or *description*, task of general data characterization, which may include deviation detection, database segmentation, clustering, associations, rules, summarization, visualization and text mining.

The process of data mining starts with data screening, cleaning and integration from different sources. Particularly in large, integrated databases where the data come from many different sources, there will likely be errors. With the assistance provided by metadata, screening systems will identify anomalous data. In most cases some data transformation will be used to prepare datasets before model running. Care should be taken during data transformation not to mask the features that carry the most important information.

Selection of training and validation datasets using appropriate sampling strategies is the next step. In order to achieve robustness and generalization, data mining is commonly done on a split dataset. The training set is used to develop the model and to evaluate fitness of the learned model, while the validation dataset is used to calculate the overall error between the modelled and target output. It is important to include a sufficient number of parameters (data fields) that may have some relevance to the problem being

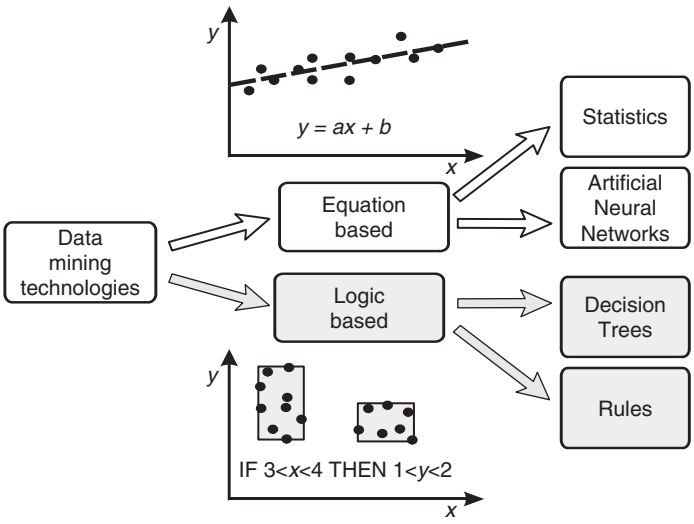

Figure 10.6 Main technologies for data mining

Source: Savic et al., 1999.

studied. The data mining system will then discover which ones are the most useful and what is the relationship among the parameters. Omitting a highly relevant parameter from analysis will cause deterioration in prediction performance of the system.

The final phase, knowledge discovery and encoding, involves running the system, validating the patterns discovered and finally encoding the results of data mining in software that can be used for prediction or classification purposes in future.

Data mining technologies can be classified into two major groups: equation based and logic based. The difference between these two approaches can be seen in Figure 10.6. The equation approach is carried out mostly on numerical data, using *statistics* and *artificial neural networks*. Typical of the statistical approaches is regression analysis (Figure 10.6). It works well for less complex sets of data, such as straight line or simple nonlinear smooth surfaces. However, transparency of the method (ability of humans to understand the equation) decreases with the complexity of the equation employed. On the other hand, the greatest advantage of artificial neural networks over other modelling techniques is their capability to model complex, non-linear processes without having to assume the form of the relationship between input and output variables.

The alternative approach, the logical approach, usually employs the conditional operators IF/THEN to represent the knowledge. The logical approach (as in Figure 10.6) is best at dealing with sharp-bounded properties of objects, although some fuzziness can be added in the condition operators. The logical approach can deal with both numeric and non-numeric data. Decision trees and rule induction are two of the most used techniques. Though similar, they differ in the way they discover information and more importantly in terms of their behaviour regarding new data items. The decision tree will use the simplest form of IF/THEN statements to represent the information and discover rules. Rule induction will use conditional relationships or so-called conditional logic (for example, 'IF it is raining, THEN it is cloudy') combined with associations

between data fields or association logic (for example, the observed fact that in 80% of cases when fault *x* occurs then fault *y* is also encountered).

Data mining is still a relatively young discipline with wide and diverse application. Because of the relative slowness of data mining algorithms, the working sample of the main database has to be small, well sampled and well-prepared. The user must consider if the data mining approach is the right one, or if instead a **s**imple **d**ata **q**uery (using SQL) would more efficiently solve the problem (Roiger and Geatz, 2002). This depends on the type of questions user wants to answer, and the type of knowledge he/she wants to discover:

- *shallow knowledge*, simple summaries (e.g. averages) or aggregates (totals) of an attribute over a selected set of cases, in which SQL is probably the most appropriate tool;
- *multidimensional knowledge*, information about the frequent occurrence of values of different attributes (known as 'Association Analysis') or simple associations in large databases (OLAP on the data cube can do this);
- *hidden knowledge*, about patterns or relationships that are not obvious, which the user can't guess prior to data mining;
- *deep knowledge*, about underlying patterns and relationships that can only be discovered using prior scientific or meta-knowledge. This is the current research frontier for data mining.

10.3.5 Self-organizing maps

A **s**elf-**o**rganizing **m**ap (SOM) can be considered as an equation-based data mining technique, which uses an artificial neural network algorithm in the unsupervised learning category. SOM is a data visualization technique invented by Kohonen (2001), to reduce dimensionality of data through use of the self-organizing neural networks. The need for such a technique is obvious, since humans have limited possibilities in the visualization of complex, high volume and multi-dimensional data sets. SOMs reduce dimensionality by producing a map which plots and thus displays the similarities of the data by grouping similar data items together.

Many fields of science have adopted the SOM as a standard analytical tool: statistics, signal processing, control theory, financial analysis, experimental physics, chemistry and medicine. The SOM solves difficult high-dimensional and nonlinear problems such as feature extraction (WEBSOM, 2005) and classification of images (PicSOM, 2001), acoustic patterns, adaptive control of robots, demodulation and error-tolerant transmission of signals in telecommunications. SOM tools are directly compatible with GIS environment where they can be used either for pure visualization purposes or together with principal components analysis for spatial identifying relationships and time series analysis. The SOM will assist integrated urban water management by making the interpretation of multi-source and multi-domain data easier for a range of urban water cycle scientists and managers.

10.3.6 Internet and grid systems

Advances in internet infrastructure (e.g. TCP/IP protocol, HTTP communication protocol used for retrieval of documents written in HTML and XML languages), allow

the rapid exchange of data and information between users. Expansion of wireless networks allows environmental monitoring stations to be simultaneously the WEB server. The user can access it from the internet and obtain status and current readings online.

Currently, the most challenging issue regarding an integrated water environment driven by internet proliferation is *web-based modelling* (Islam and Piasecki, 2004). This is the new paradigm shift in hydrodynamic modelling as well as in numerical modelling. To access the data and simulation model, the user (client) requires only a network connection, an intelligent browser and installation of required plug-ins, or client version of proprietary software. From the server side, in addition to the database management system and a high volume of disk space, a distributed version of the numerical model is needed, to share the data and tasks with one or more clients at the same time. Because of the potentially large amount of necessary data transfers between the client and server, a *client-side-request* and *server-side-simulation* approach should be used. Model-View-Controller (MVC) architecture can be used for this kind of system (Kurniawan, 2002). This separates the *simulation model* or business logic from the *model views* or presentation logics (Islam and Piasecki, 2004).

The key issue in development of such a web-based simulation environment is standardization. The development of a web markup language specific to the hydrologic community, such as HYDROML (Piasecki and Bermudez, 2003), should standardize data description and thus facilitate storage, querying, analysis, retrieval and exchange among data holding sites and end-users. The markup language differs from simple metadata. As any language, it is constructed from two components; the first is the set of *grammar rules* (syntactic structure) and the second is the *dictionary* (semantics) to provide the 'words'. The latter is derived from the metadata and associate standards, while the former is provided through the use of the Extensible Markup Language, XML (W3C-XML, 2002). While quite a number of markup languages are currently being developed and used in a variety of areas, two stand out as perhaps closest to hydrologic sciences, namely, the Geographic Markup Language (GML, 2002) developed by ISO (norm 19136) with the OpenGIS community and the Earth Science Markup Language (ESML, 2002) originating from an effort to incorporate data elements from the earth observation community.

The further extension of web-based simulation leans toward *grid systems* and *grid computing*. A grid system is 'an ambitious and exciting global effort to develop an environment in which individual users can access computers, databases and experimental facilities simply and transparently, without having to consider where those facilities are located' (RealityGrid, 2001; Joseph and Fellenstein, 2004). It offers a model for solving massive computational problems by employing the unused resources (CPU cycles and disk storage) of a large numbers of disparate, often desktop, computers treated as a virtual cluster embedded in a distributed telecommunications infrastructure. The focus is on the ability to support computation across administrative domains, which is different from traditional computer clusters or traditional, distributed computing.

In spite of rapid developments in internet-based applications, there are still a number of considerations that need to be resolved. Some of the most important are security, trust and portability.

Security of web-based simulations and grid systems remains an issue that must be better addressed by the industry and scientific community. In trusted environments, within

a department or within an open research community for example, where applications and data are not in a mission-critical or proprietary state, security is less of an issue. Beyond those safe harbours, there is a call for strong security measures, and the industry is working to overcome both real and psychological firewalls. Currently, several working groups within the Global Grid Forum are working on standardization of the many already existing security solutions.

Another important question is the degree of trust that a user can place in results and views obtained from a web-based simulation model. What form of accreditation of data is needed, and how will data uncertainty be communicated to the users? There is also a question of information availability: how to be sure that today's accessible pages and sites with relevant data and literature will be accessible tomorrow? Saving local files is a solution, but adds to storage requirements and data duplication.

Another consideration concerns portability of data. Now it is common to share data over the internet between people working on the same project, but questions still remain about how others can use the data. Data formats and accompanying metadata are tightly linked with programmes and applications that will use the data. Conversion from one format to another is sometimes painful (for example, conversion between rainfall-runoff models, say, from Hydroworks to Mouse or Simpson) or sometimes not possible without additional work (if concepts of programmes were not the same, such as in the conversion from a node-oriented model to a link-oriented one).

10.3.7 Object orientation

The Object Oriented (OO) concept is based on a real world concept: the fundamental construct is the *object* that combines both *data structure* and *behaviour* (Fedra, 2003). The OO concept was first developed as a database management system, but since then it has evolved into the mature concept capable of integrating a wide range of information technologies. Most programming languages are now object oriented, and GIS also works well within the object-oriented umbrella. In addition, databases with sharable resources and integrated metadata are becoming object oriented, and even numerical simulation models are being developed using object-oriented approaches.

In the OO system, the object is the basic element. Objects of the same structure and behaviour are grouped within classes. The basic properties of objects are (Prodanović, 1997):

- *Abstraction*, a mental process that extracts the essential aspects of the object that distinguishes it from all other kinds of objects for a particular purpose. For example, the complex real world object is reduced to the rectangular shape named 'Poly1' to represent a certain cover type, the percentage of imperviousness, specific wastewater flow, and so on. There are four basic concepts of abstraction: *classification*, *generalization*, *association* and *aggregation*.
- *Inheritance*, a mechanism permiting the development of new classes by modifying the existing one. The object 'Poly1' will have all characteristics of classes and subclasses (for example, the 'Cover' subclass used to describe the terrain cover) it was derived from. This is the primary way of minimizing data redundancy and of breaking the complex, real-world objects into manageable modules.

- *Encapsulation* or *content hiding*, which means that everything within the object is private to that object. The only way to work with its contents is through the object interface, the built-in *object operators*, *methods* and *rules*. The area of 'Poly1' is stored within this object. If the 'Cover' classes used to derive the 'Poly1' have a function that will report the polygon area, the user can ask the 'Poly1' for its area. Otherwise, access to that information is not possible.
- *Polymorphism*, meaning that the same object can respond differently, depending on the type of operation the user asks it to perform, and the current state of the object. The total area operator in the 'Cover' class will look at the 'Poly1' object to check if it is a background type of area or not, so as whether to include the area into the calculation or not.
- *Message passing*, meaning that the only way to work with an object is to pass a message to the object and to wait for the object's response. If the built-in methods in the object recognize the received message as a valid operation, they will react; otherwise they will ignore the message. The message system is very important, and it is employed for user communication with the objects, as well as internal communication among the objects.
- *Persistence*, meaning that an object will live as long as the (authorized) user decides so. Anything that refers to the object will then also delete references to it.

From the given list of an object's basic properties in OO systems, it can be seen that OO can be easily applied to distributed databases and the distributed concept of computing and modelling. A large database that covers all needs of a big water supply company can be separated into a number of smaller units (see Figure 10.3). All those units could be spread around the company, sitting on different computers within appropriate departments and using different operational systems. An intranet or full internet link can be used to connect the computers and databases. The object orientation approach is thus very helpful in achieving database integration and sharing.

There are a number of examples of the OO approach in resolving water resource management problems. Fedra and Jamieson (1996) have described and used three spatially referenced object types: river basin objects, network objects and scenarios. Those objects have functions that can obtain or update their current state and report the state to clients, and a number of classes used to derive the objects. Havnø et al. (2002) gave an excellent example of OO code architecture development on the Caloosahatchee Basin in central Florida. Another large project (WaterWare, 2005) came out as a result of EUREKA EU487. It is an integrated model-based information and decision support system for water resources management, developed and applied on the River Thames (England), the Lerma-Chapala Basin (Mexico), the West Bank and Gaza (Palestine), the Kelantan River (Malaysia) and the Yangtze River (China). Also, applications around the Mediterranean in the EU sponsored projects SMART and OPTIMA included river basins in Cyprus, Turkey, Lebanon, Jordan, Palestine, Egypt, Tunisia and Morocco. The OO approach is widely used also in Cooperative Research Centre for Catchment Hydrology [http://www.catchment.crc.org.au/ (Accessed 02 July 2007.)] for a development of catchment modelling toolkit [http://www.toolkit.net.au/ (Accessed 02 July 2007.)]. These examples provide useful 'starting points' to identify what may be possible for a specific local application.

10.4 FROM INFORMATION TO KNOWLEDGE

Using all the described techniques in the previous subsection, the user can generate a vast amount of information. However, the information will become knowledge only after the user is able to process and understand it. Knowledge itself, however, is not the final goal; the ultimate goal is to use that knowledge to make decisions.

10.4.1 Resolving data bottlenecks

Generally, it is the transfer of information into knowledge that provides the bottleneck in the overall flow from data to knowledge. The 'transfer rate' of information into knowledge is limited, and in fact, the large rate of information production can cause phenomenon known as 'information infarct' due to information overload (Maurer, 2003, 2004).

The problem of a user's limited capacity in processing acquired information is even more challenging if this is to be performed in a limited time frame, for instance, in order to undertake action in the case of emergency or disaster. In order to make significant progress towards understanding more complex integrated environments and to undertake the right actions at the right time, it is crucial to improve handling of the ever-growing amount of information. Some measures that could be considered are (Maurer, 2003):

- *Accelerating information transfer rates*.
- *Homogenization and standardization of information representation*. It is much easier for the user to work with standardized diagrams and tables, than to have to re-learn each time table headers and diagram axes.
- *Improvement of selection mechanisms for targeted information retrieval* to provide only the required information and thus reduce overload. Technology can help in selection and ordering of the most relevant information (a good example is the Google web search engine, which in the most cases will offer the most relevant information within first 5 to 10 listings).
- *Improvement of aggregation schemes to summarize information*. The pile of information can be reduced if usable aggregate information is presented to the user. Then, if users want deeper knowledge, they can further interrogate the summarized information.
- *Improvement of disaggregating schemes and interface definition* to facilitate sharing of the work-load among a number of project participants, thus decreasing the amount of information required by a single individual. Since more users will be involved in knowledge creation, the process will be quicker. Also, each user can select the most appropriate type of information according to need.

These proposed measures call for the organization of better coordinated structures with a high degree of complexity, both in a technical and an administrative sense. Examples in the technical domain include libraries, meta-databases, standardization efforts, generic concepts, expert and decision support systems. Similarly, in the human domain, this relates to improved coordination of organizations and programmes at the international, national, regional and local levels. Often developments in the technical

field trigger change in organizational structures within society. Sharing of the workload on a global scale is thus needed to cope with complex tasks that are to be solved. In the global knowledge society this leads toward establishment of balanced and standardized education systems throughout the world. Chapter 11 provides some further discussion of the institutional requirements to support integrated data collection and use for urban water management.

10.4.2 Knowledge application

The true value of knowledge is only in its use. As shown in Figure 10.2, there are two main outcomes from knowledge creation: the feedback to various parts of the system and the action (or the decision *not* to take action) that is undertaken as a result of this knowledge.

Feedback towards the start of data-information chain will in general improve the performance of the whole system. Optimization of monitoring network, sampling site position, sampling frequency and methods used for data evaluation, based on early results and previous knowledge, can significantly reduce flow of the data, but improve its quality. Improvements to models can help to understand existing data.

The feedback will be also applied to autonomous real-time controllers (RTCs), devices that mimic data to knowledge transformation by taking some actions based on obtained data. Through the process of learning about the whole system and behaviour of each RTC, the controlling strategy can be changed and adapted to optimize the systems performance.

Actions that the user will undertake based on knowledge will be the result of a *decision-making process*. Computers aid this process through **d**ecision **s**upport **s**ystems (DSS). The objective of DSSs for integrated urban water management is to improve planning and operational decision making processes by providing useful and scientifically sound information in a dedicated form to the actors involved in these processes, including public officials, planners and scientists, various interest groups, major water users and possibly even the general public.

A DSS will support and facilitate the process of assessing the possible consequences of measures and actions, before making a proper selection from the available alternatives. The ultimate objective is to ensure sufficient and sustainable water resources, thus contributing to the maximization of some (rather hypothetical) social welfare function.

Decision making involves a choice between alternatives. The DSS should help the user to analyse the alternatives and to rank them according to a number of selected criteria by which they can be compared. These criteria are checked against the objectives and constraints involving possible trade-offs between conflicting objectives. The constraints are to be checked also, if no alternative can meet them.

Approaches in DSS span a wide range of conceptual levels, such as (Fedra, 2003):

- *information systems*, to provide information about the present state of a system permitting forecasts based on the observed trends;
- *scenario analysis*, to support the exploration of numerous 'what if?' questions;
- *comparative evaluation*, to assess different scenarios using performance indicators established (preferably according to some local, national or international standard)

in at least two scenarios (with graphical display of data allowing easy comparison in most cases); and

- *optimization*, to reach a consensus. Since each scenario is described by more than one performance variable, direct comparison does not necessarily leads to clear ranking. This can be resolved by introduction of a preference structure that defines the trade-offs between objectives. Numerous optimization techniques are then used, either directly, or more often with a discrete multicriteria approach, that will seek an efficient strategy to satisfy all the actors and stakeholders involved in the water resource and environmental management decision processes.

REFERENCES

Babovic, V., Drecourt, J.P., Keijzer, M. and Hansen, P.F. 2002. A data mining approach to modelling of water supply assets. *Urban Water*, No. 4, pp. 401–14

Burrough, P.A. 1993. *Principles of Geographical Information Systems for Land Resources Assessment*. Oxford, Clarendon Press.

Drécourt, J-P. and Madsen, H. 2001. Role of domain knowledge in data-driven modeling. Paper presented at Fourth DHI Software Conference, 6–8 June 2001, Helsingør, Denmark.

ESML. 2002. Earth Science Markup Language. http://esml.itsc.uah.edu/ (Accessed 02 July 2007.)

Fedra, K. 2003. From data management to decision support system. N. B. Harmancioglu, S. D. Ozkul, O. Fistikoglu and P. Geerders (eds), *Integrated Technologies for Environmental Monitoring and Information Production*. Dordrecht, Kluwer Academic Publishers, pp. 395–410 (NATO Science Series, IV: Earth and Environmental Sciences, Vol. 23).

Fedra, K. and Jamieson, D.G. 1996. An object oriented approach to model integration: a river basin information system example. K. Kovar and H.P. Nachtnebel (eds), *HydroGIS"96: Application of Geographic Information systems in Hydrology and Water Resource Management*. Wallingford, UK, IAHS (IAHS Publication No. 235).

Fletcher, T.D. and Deletić, A. 2007. Observations Statistiques d'un Programme de Surveillance des Eaux de Ruissellement; Leçons pour l'Estimation de la Masse de Polluants [Statistical Observations of a Stormwater Monitoring Programme; Lessons for the Estimation of Pollutant Loads]. Paper presented at the NOVATECH 2007: Sixth International Conference on Sustainable Techniques and Strategies in Urban Water Management, 25–28 June 2007, Lyon, France.

GML. 2002. Geographical Markup Language. www.opengis.net/gml/ (Accessed 03 July 2007.).

Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques*. San Diego, CA, Academic Press.

Harmancioglu, N.B. 2003. Integrated data management: Where are we headed? N.B. Harmancioglu, S.D. Ozkul, O. Fistikoglu and P. Geerders (eds), *Integrated Technologies for Environmental Monitoring and Information Production*. Dordrecht, Kluwer Academic Publishers, pp. 3–16 (NATO Science Series, IV: Earth and Environmental Sciences, Vol. 23).

Havnø, K., Sørensen, H.R. and Gregersen, J.B. 2002. Integrated water resources modelling and object oriented code architecture. Copenhagen, DHI Water and Environment. http://www.dhisoftware.com/Bangkok2002/Proceedings/Papers%20Bangkok/BA%20029/code-architecture.doc (Accessed 02 July 2007.)

IBM. Data mining software built into the cyberinfrastructure system. www-306.ibm.com/software/data/iminer/ (Accessed 02 July 2007.)

IBM DB2. 2003. *Embedded Analytics in IBM DB2: Universal Database for Information on Demand*. IBM Corporation. ftp.software.ibm.com/software/data/pubs/papers/embeddedanalytics.pdf (Accessed 02 July 2007.)

Islam, A.S. and Piasecki, M. 2004. A strategy for web-based modeling of hydrodynamic processes. Paper presented at Seventeenth ASCE Engineering Mechanics Conference, 13–16 June 2004, University of Delaware, Newark.

Joseph, J. and Fellenstein, C. 2004. *Introduction to Grid Computing*. Prentice Hall.

Katopodes, N.D. 2003. Adaptive control of flow and mass transport by multi-sensor arrays. Paper presented at Thirtieth IAHR Congress, 24–29 August, Thessaloniki, Greece.

Kohonen, T. 2001. *Self-organizing Maps*, 3rd edn. Berlin, Springer (Springer Series in Information Sciences, Vol. 30).

Kurniawan, B. 2002. *Java for the Web with Servlets, JSP, and EJB*. Indianapolis, Ind., New Riders Publishing.

Maksimović, Č. 1999. *Y2K2C Project Initiative – Mission Statement*.

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

———. 2004. Transboundary and transdisciplinary environmental data and information integration – an essential prerequisite to sustainably manage the Earth System. Online Conference on INDUSTRY IDS – Water Europe 2004 [http://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, Thessaloniki, Greece.

PicSOM. 2001. Methods and systems for content-based image retrieval. Helsinki, Laboratory of Computer and Information Science, Helsinki University of Technology. http://www.cis.hut.fi/projects/cbir/ (Accessed 03 July 2007.)

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.

Raper, J. 1997. Geographic information on the web. Paper presented at ESF GISDATA Final Conference on Geographic Information Research at the Millennium, 13–17 September 1997, Le Bischenberg, France http://shef.ac.uk/uni/academic/D-H/gis/raper.html (Page now deleted.)

RealityGrid. 2001. Engineering and Physical Sciences Research Council, UK. http://www.realitygrid.org/index.shtml (Accessed 03 July 2007.)

Roiger, J.R. and Geatz, M. 2002. *Data Mining: A Tutorial Based Primer*. Addison-Wesley Publishing.

Santos, M.A. and Rodrigues, A. 2003. Information technology and environmental data management. N. B. Harmancioglu, S. D. Ozkul, O. Fistikoglu, and P. Geerders (eds), *Integrated Technologies for Environmental Monitoring and Information Production*. Dordrecht, Kluwer Academic Publishers, pp. 39–52. (NATO Science Series, IV Earth and Environmental Sciences,Vol. 23)

Savic, D.A., Davidson, J.W. and Davis, R.B. 1999. Data mining and knowledge discovery for the water industry. D. A. Savic and G. A. Walters (eds), *Water Industry Systems: Modelling and Optimisation Applications,* Vol. 2, Baldock, UK, Research Studies Press, pp. 155–64.

SDSC – San Diego Supercomputing Center. 2004. http://www.sdsc.edu/ (Accessed 03 July 2007.) and http://www.sdsc.edu/srb/Pappres/Pappres.html. (Accessed 02 July 2007.)

Shaimardanov, V.M., Mikhailov, N.N. and Vorontsov, A.A. 2003. Perspective decisions and examples on the access and exchange of data and information products using web and XML applications. N. B. Harmancioglu, S. D. Ozkul, O. Fistikoglu and P. Geerders P. (eds), *Integrated Technologies for Environmental Monitoring and Information Production*. Dordrecht, Kluwer Academic Publishers, pp. 435–48. (NATO Science Series, IV Earth and Environmental Sciences, Vol. 23).

Sholten, H.J. and LoCascio, A. 1997. GIS application research: history, trends and development. Paper presented to ESF GISDATA Final Conference on Geographic Information Research at the Millenium, 13–17 September 1997, Le Bischenberg, France http://shef.ac.uk/uni/academic/D-H/gis/key3.html (Page now deleted.)

Singh, V.P., Strupczewski, W.G. and Weglarczyk, S. 2003. Uncertainty in environmental analysis. N.B. Harmancioglu, S.D. Ozkul, O. Fistikoglu, and P. Geerders (eds), *Integrated Technologies for Environmental Monitoring and Information Production*. Dordrecht, Kluwer Academic Publishers, pp. 141–58. (NATO Science Series, IV: Earth and Environmental Sciences, Vol. 23)

Thearling, K. 2007. Data Mining Tutorial. http://www.thearling.com/dmintro/dmintro.html (Accessed 03 July 2007.)

W3C-XML. 2002. World Wide Web Consortium (W3C). http://www.w3.org (Accessed 02 July 2007.)

Wachowicz, M. 2000. How can knowledge discover methods uncover spatial-temporal patterns in environmental data? Paper presented at SPIE Aerosense 2000 Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology II, 24–28 April 2000, Orlando, Florida, USA.

WaterWare. 2005. A Water Resources Management Information System. http://www.ess.co.at/WATERWARE/ (Accessed 02 July 2007.)

WEBSOM. 2005. Self-Organizing Maps for Internet Exploration http://websom.hut.fi/websom/ (Accessed 02 July 2007.)