A GIS-based tool for distribution system data integration and analysis

Martin Trépanier, Vincent Gauthier, Marie-Claude Besner and Michèle Prévost

ABSTRACT

The causes of water quality problems in distribution systems are difficult to identify because they can be related to numerous sources. A tool has been developed to integrate and analyse water distribution system data with the help of geographical information system (GIS) technologies. This approach uses a flexible software architecture to gather data on distribution system structural elements, water quality sampling and especially distribution system events, all of which can be key to explaining water quality problems. The tool has been applied to five water utilities in North America and Europe, all with different data formats and data gathering practices. The approach was successful in explaining about 40% of positive coliform samples at the Laval (Quebec) utility. It also led to better data quality and responsiveness at the utilities.

Key words | Distribution system, geographic information systems, integrated approach, water quality

Martin Trépanier (corresponding author)
Mathematics and Industrial Engineering Department,
École Polytechnique de Montréal,
PO Box 6079, Station Centre-Ville, Montréal Québec,
H3C 3A7 Canada
Tel: +1 514 340 4711 X4911
Fax: +1 514 340 4173
E-mail: *mtrepanier@polymtl.ca*

Vincent Gauthier
Veolia Water – Générale des Eaux,
Direction technique,
103 rue aux Arènes, BP 60045 57003 Metz,
France

Marie-Claude Besner
Michèle Prévost
Civil, Geological and Mining Engineering Department,
École Polytechnique de Montréal,
PO Box 6079, Station Centre-Ville, Montréal Québec,
H3C 3A7 Canada

INTRODUCTION

Water distribution systems (DS) involve a diversity of structural components, water flow characteristics and environmental aspects. Therefore, computer tools for effective management of a distribution system should be capable of integrating data related to hydraulics, water quality and operational events (breakages, complaints) in order to gain a better picture of situations occurring in the system. Unfortunately, although there are many tools available for hydraulic modelling, only a few of them are designed to integrate the operational and quality data that are needed to successfully analyse specific spatio-temporal events such as water quality problems. Some centralised systems and solutions do exist, but they demand a large amount of resources, both human and material, which are not always available to utilities. There are also some difficulties to cope with uncertain, missing and erroneous data.

This paper presents the development of a knowledge-representation GIS-based tool for distribution system data integration and analysis. The tool is part of an "integrated approach", proposed for the explanation and prevention of coliform events in water distribution systems (Besner *et al.* 2001). The application of this integrated approach was subsequently further tested in five municipalities in Europe and North America (Montreal, Laval and Moncton, Canada; Egham, UK; Caen, France) to evaluate its applicability to different utilities and its performances in explaining several types of water quality problems.

This spectrum of applications leads to a variety of forms and quantities of data, resulting in a GIS-based approach which has to be more flexible than traditional application-centric software. In this paper, data integration will be addressed as the most important topic because of the

doi: 10.2166/jh.2006.004

challenges that it presents in operating a utility. The concepts of ontology and knowledge representation must be looked at in order to understand the usability of the methodology. Then, the architecture and methodological aspects of the integrated approach will be presented. One of the characteristics of the approach is that it uses a lightweight computer architecture, which is not costly and at the same time is flexible enough to be compatible with the computer environments of the various utilities. The paper will then demonstrate the ability of the tool to integrate multiple sources of data from different actors within the utilities and provides some results obtained from the exercise.

DATA INTEGRATION: A BROAD CHALLENGE

Bringing together data on water network structural elements, events taking place in the systems and water quality into a single integrated application is not a simple task, since different parameters are involved in each case. As reported by Ray (1996), water utilities are usually asset-driven, and distribution system management functions are divided among several departments. Typically, the water network is managed through the usage of hydraulic model simulation software in the engineering department. The assets inventory is usually the responsibility of the public works department, where there is a concern about the age of the network elements and their maintenance. Finally, water quality data are gathered by more "quality sensitive" teams generally located at the treatment plant. There, the concerns are related primarily to microbiological quality and measures such as pH, disinfectant concentration, turbidity and other quality features. Into this sparse distribution of tasks we can add customer complaints compilation, hydrant operation by fire fighting crews, street cleaning and a series of other actions which can take place in the distribution system and which are performed by other city or regional organisations.

There is a need to integrate data from these various sources, especially to examine water quality problems that can bring systems out of compliance with respect to regulations and public health, such as the occurrence of coliforms in distributed water. As a matter of fact, there can be several causes for coliform events, as listed by Besner *et al.* (2002a): plant breakthrough, local contamination (intrusion) following pipe breaks, survival or regrowth of micro-organisms, network deterioration, pipe flushing, etc. Data on each of these elements should therefore be gathered in order to conduct the best possible analysis. For example, a bacterial regrowth model has been examined by Zhang *et al.* (2004) to understand the interactions between these parameters but no real world experiments were conducted. Gauthier *et al.* (2001) identified five categories of topics on which data should be collected to improve the understanding of water quality variations: water quality measures, characteristics of DS, hydraulics, DS events and external events (weather). Let us remind ourselves that, as reported by Beck & Lin (2003), the main difficulty is to cope with the complexity of the interactions between the measures in water quality, in addition to the difficulty in coping with missing and erroneous data.

Interesting efforts have been expended to develop integrated DS network management systems. Boulos *et al.* (1997) proposed a database-centric approach using AutoCAD where data are gathered on several layers of information about network structure (pipes, valves), complaints, fire hydrants and other relevant components of the DS. Here, data are directly input by the system and so the information structure has to be known from the start. Some authors have proposed GIS-centred applications (McCorley 2000) in which information has to be spatialised. Barcellos (2000) superimposed data on census, water quality, DS and health in a GIS environment and tried to find relationships between them. GIS was also coupled to spatial system dynamics for simulation of water resource systems by Ahmad & Simonovic (2004). Chan & Valmores (2003) proposed an intranet-based filing system for water distribution planning documents and Kamojjala *et al.* (2003) presented interesting works on the integration of models, enterprise, SCADA and spatial data for planning and operational support, but with few references to water quality data. Finally, the vision of Hinthorn *et al.* (2003) about real-time monitoring of water quality in distribution systems should be encouraged, but for now water utilities do not all have the resources needed for this kind of system.

Technical problems arise when attempting to group data from different sources and in different formats, when no assumptions are made about their quality or structure.

For example, structural data can be obtained from hydraulic modelling software, in a given set of $X-Y$ coordinates and with given attributes which may be incompatible with the GIS. Quality data can be stored in non-geocoded, spreadsheet files. Complaints data can be paper-based. Definitions and categories of complaints, events and pipe breaks can vary from one utility to another, or even between departments in the same utility. There can be other problems as well, among them: different measurement units, coordinates which may not match (geographic imprecision), different languages, etc.

Once data have been correctly brought together, they must then be adequately analysed. Savic & Walters (1999) identified several technologies related to hydroinformatics which could be used for data analysis: data mining and knowledge discovery, geographic information systems, artificial neural networks and genetic algorithms.

ONTOLOGY

Ontological approaches have proven to be effective in domains such as speech recognition, grammar editing and semantic webs (Bouquet *et al.* 2004). Similar methods can be applied to water quality data analysis. We emphasise here the importance of knowledge representation, the use of an object-oriented approach and the accountancy of the level of resolution in the analysis process.

Knowledge representation

"Knowledge representation (KR) is the study of how knowledge about the world can be represented and what kinds of reasoning can be done with that knowledge." This short definition by Ginsberg (1993) summarises the main design challenges for any software intended for DS analysis. To achieve an adequate knowledge representation of water quality events in a distribution system, the proposed approach recognises the following expert-system basics: 1) there is a trigger event to analyse, 2) inference rules are to be derived in order to search for the cause of the event, 3) the rules are applied to data sets and the results are observed with the most appropriate tools (GIS, spreadsheet) and 4) feedback from inference rules is used to refine the criteria.

In order to apply an adequate knowledge representation, it is important to integrate data in such a way that the application is not influenced by data structure. Data structure should be indulged, not overcome (Davis *et al.* 1993). Therefore, more data does not necessarily mean better analysis if the data have not been properly integrated. A phenomenon, which could be called "data indigestion", can occur when there is more data available than the system can absorb. For example, one automated data collection system would provide several thousand records in a single day, while other available data are manually gathered once a day. To avoid data indigestion, there is a need to focus on central data pieces with specialised querying features, at an adequate level of resolution.

Object-oriented approach

Object-oriented approaches have succeeded in organising and analysing large quantities of data (Trépanier & Chapleau, 2001). The identification of objects in a system helps to define their roles and the events that can be triggered by them, thus facilitating computer programming of applications. Object-oriented techniques were successfully implemented by Spanou & Chen (2001) in the case of the Upper Mersey River water quality modelling.

A broader image of a system can be obtained, then, by establishing relations between objects. In the present case, this object-oriented approach is separated from the programming language and database storage in order to avoid any influence from one to the other. The object model in Figure 1 presents a non-exhaustive list of objects which are involved in the "life" of a distribution system. Any object could trigger a water quality event and therefore must be considered in the object model. For example, contractors could cause a microbial intrusion event during a pipe repair, land use change could cause a variation in water demand, etc.

Objects are usually easier to identify than events. Before gathering data, some work must be carried out to classify and characterise objects and events in order to arrive at a uniform basis for the comparison of utilities. Complaints are a good example of ambiguous event objects. At one utility, phone complaints from customers were distinguished from information calls related to invoices and accounts, for example,

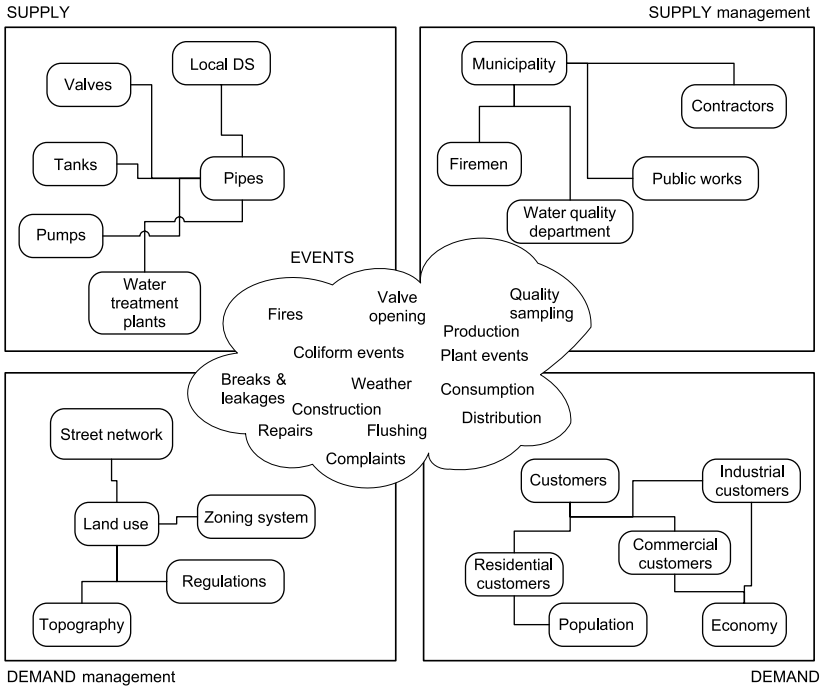


Figure 1 | Example of object model for the Distribution System.

while this was not the case elsewhere. Work has been done to find uniform definitions, if needed.

Level of resolution

The level of resolution is another factor to be considered when collecting data from different sources. In a traditional GIS-based application, the spatial levels of resolution of data have to be similar in order to superimpose information in layers. In this project, many differences were encountered, such as: complaints registered at address level, pipe breaks coded at street level, valve operations for pipe flushing entered at zone level and events without geographic information. Units and projections of all kinds were also found: longitude/latitude, in-house coordinates and hydraulic software coordinates.

The temporal level of the resolution of data must also be examined. Figure 2 shows a timeline where a water quality problem (coliform event) is illustrated with a vertical line at time "*t*". Sampling data are punctual events not necessarily synchronised with the "real" water quality (coliform) event, at the specific time it occurs. Other events, like complaints, may have an assigned file date (i.e. the date of the phone call), but are associated with DS events which can have occurred many hours or days before. These timescale differences must be taken into account when performing database queries to investigate the causes of water quality events.

The problems related to the level of resolution of data must be addressed with great care. It is important not to lose pieces of information while trying to fit data to the spatial and temporal units chosen.

METHODOLOGY

This section describes the methodology that has been found to be the most suitable for data integration and analysis, regarding the data obtained from the five utilities. At first, the need for a GIS-based tool was clearly defined, since spatial and temporal proximity must be considered when querying water quality events.

Architecture

The architecture of the GIS-based tool is lightweight and open. It integrates four software components, as shown in the data flow diagram in Figure 3. Data from water utilities

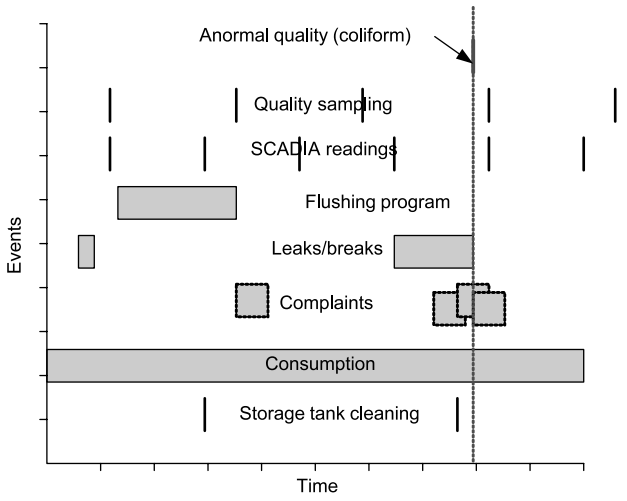


Figure 2 | Temporal level of resolution of DS data.

are collected and geocoded, and then put into Excel workbooks, accompanied by metadata. DS hydraulic modelling is performed with EPANET software version 2.00.10 (Rossman 2000). Information is consolidated by means of the in-house software, IMADSIG (in French: Interface de modélisation et d'acquisiton de données pour un système d'information géographique (data modelling and acquisition interface for a geographic information system)). With this software, users have the ability to identify various types of quality events to be analysed and to create circumstantial queries. Data are then extracted and converted into GIS format for visualisation in the GIS data viewer ArcExplorer (ESRI, Redlands, CA).

Water utilities use other tools as well. Therefore, data must be exchanged between these tools and this approach. The City of Laval uses the 4D database management system for some data gathering (4D Inc., San Jose, CA). Therefore, an interface has been put in place to directly transfer output into Excel files (Microsoft Corp., Redmond, WA). Other software, such as Microstation (Bentley Systems, Exton, PA) and AutoCAD (Autodesk, San Rafael, CA), were also involved in our work, showing that the open architecture is quite flexible.

Data gathering

To facilitate exchanges and manipulation, a spreadsheet application (Excel) is used to gather and structure data. Excel is a widely used spreadsheet application with a compatible file format in its three most recent versions (XP, 2000, 97). This medium has proved to be suitable for this project and its multiple data formats, since many distribution system operators are "spreadsheet-literate". Workbooks are created to store the data. Typically, they include data on a major topic, with several worksheets as subtopics. For example, the "Network Structure" workbook contains "Nodes" and "Links" worksheets.

To document data structure and usage, a metadata workbook is used to describe the contents and the use of all data workbooks and their sheets (Figure 4). In addition to the file list, metadata includes file properties such as: descriptions, dates, user, geographic information

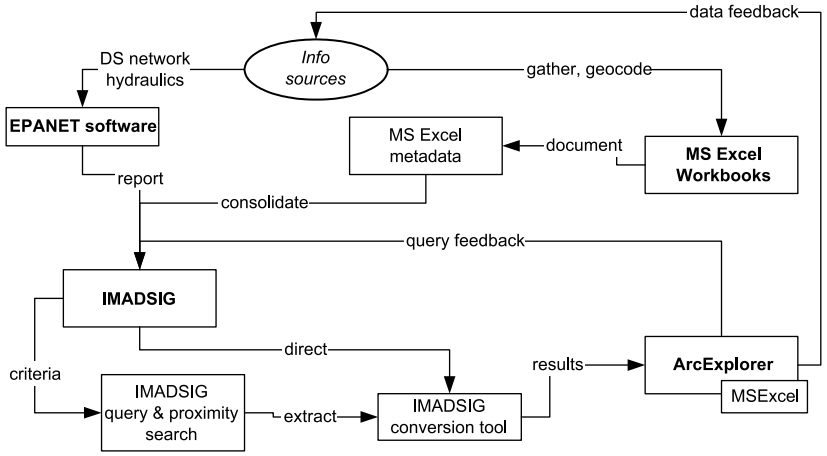
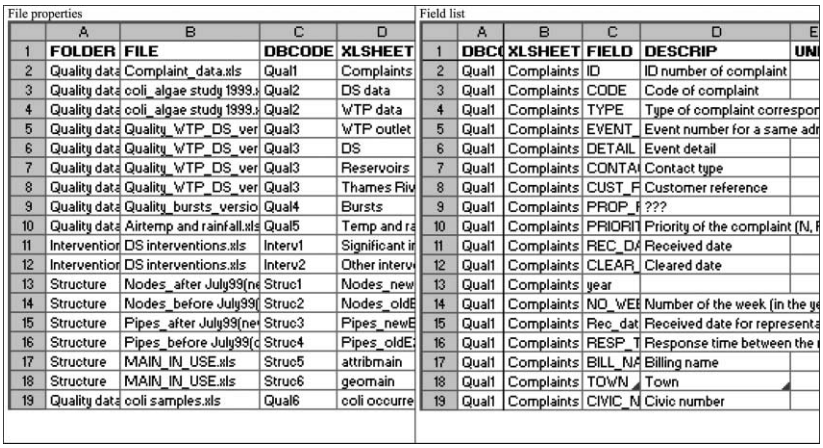


Figure 3 | Data flow diagram for the GIS-based tool.



File properties

	A	B	C	D
1	FOLDER	FILE	DBCODE	XLSHEET
2	Quality data	Complaint_data.xls	Qual1	Complaints
3	Quality data	coli_algae study 1999.x	Qual2	DS data
4	Quality data	coli_algae study 1999.x	Qual2	WTP data
5	Quality data	Quality_WTP_DS_ver	Qual3	WTP outlet
6	Quality data	Quality_WTP_DS_ver	Qual3	DS
7	Quality data	Quality_WTP_DS_ver	Qual3	Reservoirs
8	Quality data	Quality_WTP_DS_ver	Qual3	Thames Riv
9	Quality data	Quality_bursts_versio	Qual4	Bursts
10	Quality data	Airtemp and rainfall.xls	Qual5	Temp and ra
11	Intervention	DS interventions.xls	Interv1	Significant in
12	Intervention	DS interventions.xls	Interv2	Other interv
13	Structure	Nodes_after July99(ne	Struc1	Nodes_new
14	Structure	Nodes_before July99(	Struc2	Nodes_oldE
15	Structure	Pipes_after July99(ne	Struc3	Pipes_newE
16	Structure	Pipes_before July99(c	Struc4	Pipes_oldE
17	Structure	MAIN_IN_USE.xls	Struc5	attribmain
18	Structure	MAIN_IN_USE.xls	Struc6	geomain
19	Quality data	coli samples.xls	Qual6	coli occurre

Field list

	A	B	C	D	E
1	DBC	XLSHEET	FIELD	DESCRIP	UN
2	Qual1	Complaints	ID	ID number of complaint	
3	Qual1	Complaints	CODE	Code of complaint	
4	Qual1	Complaints	TYPE	Type of complaint correspor	
5	Qual1	Complaints	EVENT_	Event number for a same adr	
6	Qual1	Complaints	DETAIL	Event detail	
7	Qual1	Complaints	CONTA	Contact type	
8	Qual1	Complaints	CUST_R	Customer reference	
9	Qual1	Complaints	PROP_T	???	
10	Qual1	Complaints	PRIORIT	Priority of the complaint (N,	
11	Qual1	Complaints	REC_DA	Received date	
12	Qual1	Complaints	CLEAR_	Cleared date	
13	Qual1	Complaints	year		
14	Qual1	Complaints	NO_WEE	Number of the week (in the ye	
15	Qual1	Complaints	Rec_dat	Received date for representa	
16	Qual1	Complaints	RESP_T	Response time between the	
17	Qual1	Complaints	BILL_NA	Billing name	
18	Qual1	Complaints	TOWN_	Town	
19	Qual1	Complaints	CIVIC_N	Civic number	

Figure 4 | Metadata workbook contents.

fields, GIS topology types and GIS topology fieldnames. It also contains a list of all fields from all tables (sheets) along with their description, type and size. Information on GIS visualisation attributes such as style, colour and size can also be stored. An advantage of this set-up is that field names can be changed to adapt to each system, especially in this study where both French and English names are used. Field formats and sizes can also be easily changed by users as well, with minimal constraints as to their contents and to their use of attributes.

Data processing

Some work is needed to prepare data before the IMADSIG interface software and its post-processing units are used for investigating databases. Depending on the data source and format, the relevant steps are:

- *Capture.* For some utilities, data were only available on maps or in paper-based documents. Data was manually entered in Excel and formats were transmitted back to utilities to provide them with a base on which they could continue entering data. In more advanced utilities, systematic procedures have been developed to permit direct geocoding of data with the help of an address-point file containing all the civic addresses connected to the DS with their $X-Y$ coordinate attributes.
- *Formatting.* Often, even thought data are available in numerical format, they have to be formatted for introduction into database-like software. For that reason, some data were normalised and tips were provided to utilities to help them enter normalised data. Typographical and other errors were also found and corrected at this point.
- *Validation.* Univariate and multivariate validation were necessary in order to make the data comparable and rankable. Univariate validation consists of classifying complaint types, breakage causes and other attributes. For example, in the city of Laval, complaint types were reduced from 24 to 7 in order to better categorise them. Multivariate validation was used in particular for data geocoding, where more than one field can be used for treatment. Intersection information, for example, which is composed of two street name fields, was validated in accordance with intersection files, thus simplifying the geocoding process. Address information, which is composed of a civic number and a street name, was corrected in a similar way.
- *Geocoding.* Data geocoding is an important issue to consider when gathering distribution system data with this approach. Even though some of the data may already have been geocoded by utilities, coordinate conversion and projection stretching may be needed to fit projection systems. The level of resolution of the geocoding step depends on the quality of the data and the quantity of information available. Some choices were made to match each type of data to the best topological object (Table 1). A hierarchy of geocoding levels of resolution is used to geocode incomplete information: Global Positioning Systems (GPS) $X-Y$ coordinates, if available, then civic

Table 1 | GIS topological features used for data visualisation

Feature	Usage
Point	Sampling stations, complaints (customer's residence), hydrants, valves, treatment plants, tanks, pipe breaks
Link (two points polyline)	Pipes, pumps
Polyline	Streets
Area (polygon)	Flushing zones, districts, municipalities
Attached feature	Quality measures (attached to sampling stations), treatment plant events, flushing operations (attached to flushing zones)

number (address), intersection (two street names), street arc (street name, only for short streets) and zone.

Data conversion

The IMADSIG prototype interface software was specifically developed for the project (Figure 5). The software, programmed in Visual FoxPro (Microsoft Corp., Redmond, WA), has several features: data conversion, GIS shapefile creation, hydraulic software report processing and query functions.

IMADSIG's first task is to convert XLS data to DBF (database file) format, using the information provided in the metadata workbook. The DBF format is a widely used PC-database format and is compatible with a great deal of software. Data structure consistency (good field type, field sizes, rounding) is checked during this operation. IMADSIG then converts DBFs to shapefiles. IMADSIG uses the metadata workbook to make topologic objects out of DBF flatfiles. This whole operation is called "rough" conversion and is aimed at spot-checking for errors in the files such as bad formatting, mismatched field names or types or spatial attributes that are wrong. Users can easily go back to datafiles and correct the errors as needed.

Querying databases

IMADSIG software is also used to query data in order to find causalities between DS events, DS structure, hydraulic features, quality data and water quality problems, but can also query other pieces of information such as, for example, all the pipes of a certain age, etc. Since each object datafile has its own structure and attributes, a two-part flexible querying interface was added to the tool. The first part consists of a multi-table querying wizard, where users can quickly specify queries based on spatial proximity (radius search) and temporal proximity (on or about a date) to be performed on several tables at a time. The wizard directly generates the Structured Query Language (SQL) phrases associated with the queries, which is a big advantage for people who are not knowledgeable about SQL. Multiple queries are grouped into searches which can be identified, stored and annotated by users, providing a simultaneous overview of a given situation. Search results are available through Excel and DBF files, and shapefiles are created to be visualised with a GIS viewer. A search log is provided to list the number of elements found or to report errors in the process.

Hydraulics

A special IMADSIG module analyses the EPANET simulation reports in order to reproduce the state of the distribution system during an extended simulation period to obtain stabilised results, taking into account the demand variations and various conditions of operation. From these records of several thousands of results, the Visual FoxPro database engine is used to calculate statistics such as minimal, maximum and average pressure, flow, velocity and water age for each pipe of the network. The information is then reintroduced as a data component which can be queried and displayed geographically.

This result is also used by a network proximity algorithm which has been developed to retrieve the trace of the water in the distribution system, from treatment plant(s) to a given location in the network where a water quality variation has been identified. This algorithm identifies all the pipes in which the water may have circulated to supply the specific point, in order to maximise the area where the possible source of the problem may be looked for. This quite intuitive method is presented in Figure 6 where node N1 is a specific location to be investigated.

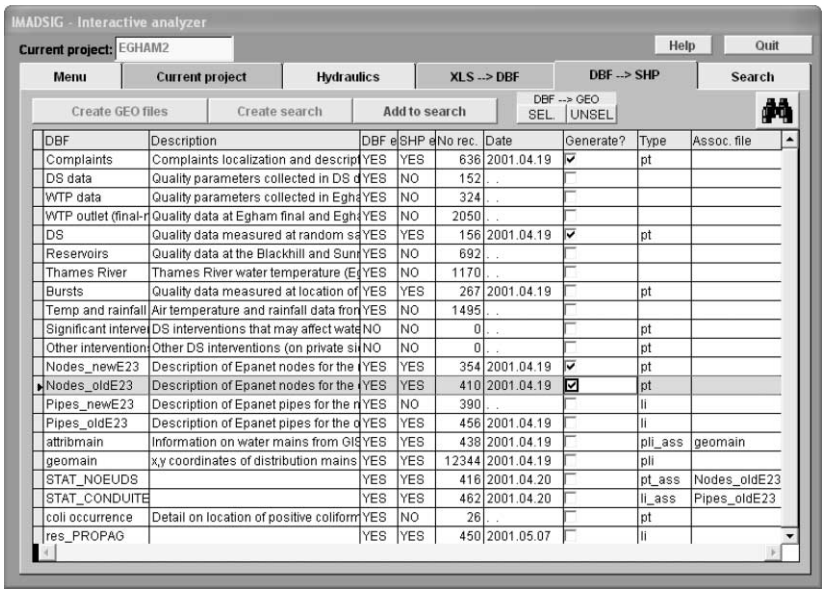


Figure 5 | IMADSIG interface for data conversion.

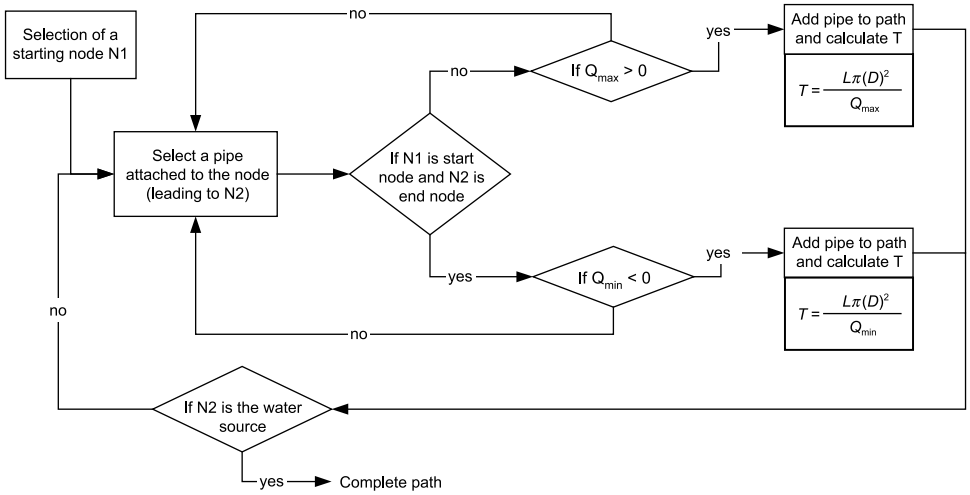


Figure 6 | Algorithm to retrieve all water paths towards a given node (T = water age, L = pipe length, D = pipe diameter, Q = flow).

GIS visualisation

This task requires the use of a GIS viewing tool. ArcExplorer, free software provided by ESRI, is powerful enough to view large geographical files and is used in this project to examine the shapefiles created by IMADSIG (Figure 7). To facilitate search result visualisation, ArcExplorer project files are produced by IMADSIG. These project files reference and format the layers created by IMADSIG during the querying process.

IMPLEMENTATION

Refinement of the methodology used for this research project is the result of a partnership between five utilities: Montreal, Laval and Moncton (Canada), Egham (United Kingdom) and Caen (France) (for Caen, please refer to Jaeger *et al.* (2002)). The City of Laval implementation of the methodology is presented here as an example of the applicability of the approach.

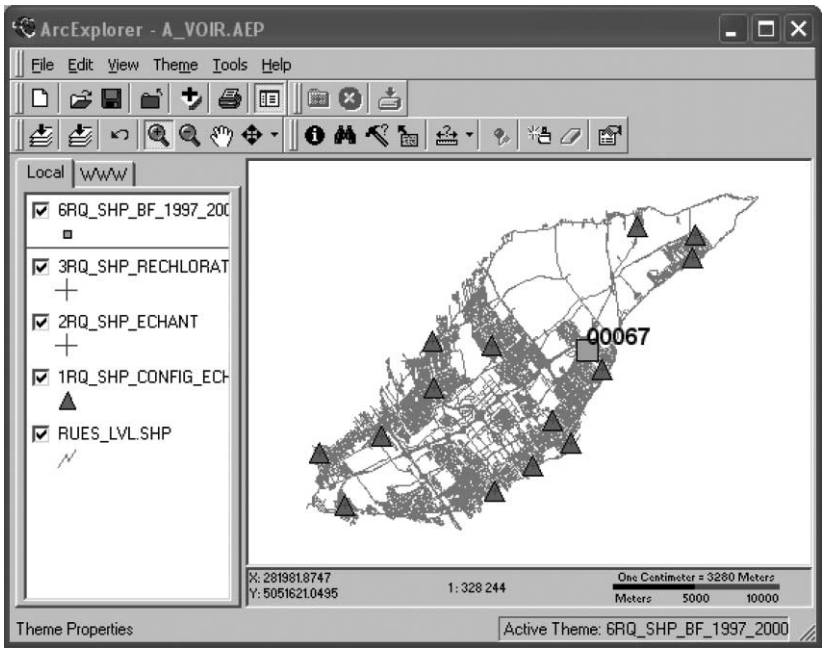


Figure 7 | Result visualisation in ArcExplorer.

Laval is a city of about 350,000 inhabitants located north of Montreal. Its 1465-km-long water distribution system is supplied by three water treatment plants. In Table 2, the Laval database gathered through this project is described. A major challenge was the geocoding step, mainly because of the large amount of data. As shown in the table, most of the data was geocoded manually. Today, geocoding is semi-automatic, thanks to the availability of an address database.

The use of the GIS-based tool in Laval was aimed at identifying the causes of water quality events, especially the occurrence of positive coliform samples. The interactive tool has been used to query the database on 88 positive total coliform samples during the time period 1997–2000. Data on treatment plant activities, water quality parameters and distribution system events were analysed. This study was the scope of a paper by Besner *et al.* (2002b). Results of the application showed that, in some cases, distribution system events such as unidirectional flushing performed as part of established programs, valve and hydrant manipulations and lower pressures may have led to microbiological water quality problems. A highly probable cause has been identified for about 43% of the positive coliform samples investigated.

DISCUSSION

The approach presented in this paper differs from other valuable work published on the subject in that five different utilities were involved in the study, all with specific needs, data availability and data formats. This explains the relative lightweight nature of the tools in comparison with some other (more costly) advanced commercial systems. Another reason for the choices that were made is the presence, in these utilities, of established procedures and software for data collection and hydraulic modelling. On the one hand, our approach had to be flexible enough to interface with these tools, while on the other, it had to accept manually entered data without much manipulation. The approach is not intended to impose a specific data standard that would have to be respected. In our view, data standards have a negative effect, since the restrictions that they impose tend to reduce the amount of usable information, with the risk of overriding some regional aspects of data formats and contents. The choice of software elements (Excel spreadsheet, ArcView GIS, EPANET and an in-house interface) reflects the openness of the approach.

This work has been marked by a paradigm shift: while trying to gather data to find the causes of the presence of

Table 2 | Available data from City of Laval. (Geocoding method: A = automatic, M = manual.)

	No. of records	Native format	Temporal coverage	Geocoding	Geocoding method
Water quality (DS and treatment plant)					
Sampling stations + plants	15	Excel	1997–2000	100%	M
Total coliform	9946	4D	1997–2000	Ass. to sampling stations and treatment plants	
Atypical bacteria	9613	4D	1997–2000	‖	
Free chlorine	3345	4D	1997–2000	‖	
Water temperature	9321	4D	1997–2000	‖	
Turbidity	9235	4D	1997–2000	‖	
Conductivity	9110	4D	1997–2000	‖	
pH (treatment plant only)	3284	4D	1997–2000	Ass. to treatment plants	
HPC (treatment plant only)	3284	4D	1997–2000		
Hydraulics					
Pipes	24303	Piccolo (Safege, France)		Data conversion	M
Nodes	21538	Piccolo (Safege, France)		Data conversion	M
Hydraulic results	70MB	EPANET/text file		Data conversion	A
Operation and maintenance					
Valve operations	32798	Excel	1997–2000	99.9%	A + M
Hydrant operations	25021	Excel	1997–2000	99.9%	A + M
Pipe breaks	3302	Excel	1996–2000	94.2%	A + M
New pipes	183	Excel	1997–2000	100%	M
Pipe rehabilitations	197	Excel	1997–2000	100%	M
Flushing zones	51	Paper maps		Manually	M
Customer complaints					
Complaint log	1951	4D	1991–2000	99.7%	A + M
Others					
Plant events		Hand-written		At plant	M
Street coverage		GIS		Already done	
Basemap		GIS		Already done	

coliforms in water distribution networks, it was found that data integration problems had to be solved first. These problems led us to consider that a special kind of tool was needed to address them. This is why knowledge representation should be used in the tool to clearly identify the objects present and the relationships between them. The use of GIS has been an obvious choice from the beginning, since problems in distribution systems are geographically distributed.

The approach also has some limitations:

- The variety of data sources and formats resulted in extra work on data validation and conversion. For example, with computer applications in five cities, two languages and on two continents, problems arose with regional Microsoft Windows settings, the decimal separator, etc.
- The absence of precise adapted event definitions in some utilities forced the use of knowledge representation in the process, thereby extending data collection delays. In some utilities, these changes had to be implemented before extra data were collected.
- The majority of data were not geocoded, and so extensive efforts were made to add geocoding procedures at the collection source.
- There are limitations due to the use of Excel spreadsheets for data gathering (65,535 records), but this can be overcome with the direct conversion of database management systems to DBF formats.

It is clear that, with this approach, utilities will have to be more "responsible" for their data collection procedures, to ensure better use of data. For example, IMADSIG led the City of Laval to add a geographic component to their own 4D system and to enter data more quickly when events occur on the network. Technicians are now more aware of data quality because they can visualise and better analyse them in an operational context. The tool is now also used in ways unintended initially, such as infrastructure maintenance and network mapping.

Some developments are still to come regarding the IMADSIG system and its components. Data exchange between corporate systems and the tool could benefit from technologies like XML (extended mark-up language). In its actual state of development, the approach has also been tested in two water utilities in the United States (Greater Cincinnati Water Works and Denver Water Department) to illustrate how data integration can help in identifying causes of customer complaints (Besner *et al.* 2003; Martel *et al.* 2005). It is also currently being applied in one French distribution system operated by Veolia Water (City of Metz).

CONCLUSION

Integrating data on distribution systems is a task that requires the strong commitment of several actors in a water utility. The GIS-based tool developed in this project, with its lightweight architecture, represents a flexible approach which has been implemented in five utilities in North America and Europe, each with its own problems and specificities regarding the water system and its database-related management. What was initially a tool to specifically reveal the causes of water quality problems has become a tool for integrating many kinds of other information on DS and is quickly demonstrating its usefulness for general DS management purposes.

ACKNOWLEDGEMENTS

This work was performed in the framework of a larger research project and was supported by Veolia Water and the Industrial Partners of the NSERC (Natural Sciences and Engineering Research Council of Canada) Industrial Chair on Drinking Water of the Ecole Polytechnique de Montréal. The Cities of Laval and Montreal (Quebec, Canada), the City of Moncton (New Brunswick, Canada), the City of Egham (United Kingdom) and the City of Caen (France), provided expertise and databases for case studies. Thanks also to Robert Chapleau for his input in the project.

REFERENCES

Ahmad, S. & Simonovic, S. P. 2004 Spatial system dynamics: new approach for simulation of water resources systems. *J. Comput. Civil Engng.* **18** (4), 331–340.

Barcellos, C. 2000 Health risk analysis of the Rio de Janeiro water supply using Geographic Information Systems. *Interdisciplinary*

Perspectives on Drinking Water Risk Assessment and Management. IAHS Publication No. 260, pp 95–99.

Beck, M. B. & Lin, Z. 2003 Transforming data into information. *Wat. Sci. Technol.* **47** (2), 43–51.

Besner, M. C., Carrière, A., Prévost, M. & Martel, K. D. 2003 Use of data integration to identify the causes of customer complaints in distribution systems. *Proc. American Water Works Association, Water Quality and Technology Conference, Philadelphia, PA, CD-ROM*.

Besner, M. C., Gauthier, V., Servais, P. & Camper, A. 2002a Explaining the occurrence of coliforms in distribution systems. *J. AWWA* **94** (8), 95–109.

Besner, M. C., Gauthier, V., Morissette, C. & Prévost, M. 2002b Identification of the main causes of total coliforms in a distribution system. *Proc. American Water Works Association, Water Quality and Technology Conference, Seattle, WA, CD-ROM*.

Besner, M.-C., Gauthier, V., Barbeau, B., Millette, R., Chapleau, R. & Prévost, M. 2001 Understanding distribution system water quality. *J. AWWA* **93** (7), 101–114.

Boulos, P. F., Heath, E. J., Feinberg, D. H. & Ro, J.-J. 1997 Water distribution network management: a fully integrated approach. *J. NEWWA* **111** (6), 180–188.

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L. & Stuckenschmidt, H. 2004 Contextualizing ontologies. *Web Semantics: Science, Services and Agents on the World Wide Web* **1** (4), 325–343.

Chan, C. D. & Valmores, M. A. 2003 Web-based GIS with a planning database. *Proc. American Water Works Association, Information Management and Technology Conference, Santa Clara, CA, CD-ROM*.

Davis, R., Shrobe, H. & Szolovits, P. 1993 What is a knowledge representation? *AI Mag.* **14** (1), 17–33.

Gauthier, V., Besner, M.-C., Trépanier, M., Millette, R. & Prévost, M. 2001 Tracking the source for microbial contamination in distribution systems using an integrated approach. *Proc. American Water Works Association, Water Quality and Technology Conference, Nashville, TN, CD-ROM*.

Ginsberg, M. 1993 *Essentials of Artificial Intelligence*. Morgan Kaufmann. San Mateo, CA.

Hinthorn, R., Wilson, L., Moshavegh, F. & Yadav, S. 2003 A vision for real-time monitoring and modelling of water in water distribution systems. *Proc. American Water Works Association, Information Management and Technology Conference, Santa Clara, CA, CD-ROM*.

Jaeger, Y., Gauthier, V., Besner, M. C., Viret, B., Toulorge, R., Lemaire, E., De Roubin, M. R. & Gagnon, J. L. 2002 An integrated approach to assess the causes of drinking water quality failures in the distribution system of Caen. *Wat. Supply Wat. Sci. Technol.* **2** (3), 243–250.

Kamojjala, S., Fang, M. & Jacobsen, L. J. 2003 Integration of models, enterprise, SCADA and spatial data for planning and operational support. *Proc. American Water Works Association, Information Management and Technology Conference, Santa Clara, CA, CD-ROM*.

Martel, K., Hanson, A., Kirmeyer, G. J., Besner, M. C., Carrière, A., Prévost, M., Lynggaard-Jensen, A. & Bazzurro, N. 2005 *Data Integration for Water Quality Management. AwwaRF Rep.* AWWA. Denver, CO.

McCorley, S. 2000 GIS application at Bergen Water. *Wat. Supply* **18** (4), 24–30.

RAY, C. F. 1996 The use of GIS in a major water utility company. *Proceedings, ICE: Civil Engineering*, 1996. 114/special issue 2, pp 23–29.

Rossman, L. A. 2000 *EPANET User's Manual*. US Environmental Protection Agency, Washington, DC. http://www.epa.gov/ORD/NRMRL/wswrd/epanet.html

Savic, D. A. & Walters, G. A. 1999 Hydroinformatics, data mining and maintenance of UK water networks. *Anti-Corrosion Meth. Mater.* **46** (6), 415–425.

Spanou, M. & Chen, D. 2001 Modelling of the Upper Mersey River system using object-oriented tools. *J. Hydroinf.* **3** (3), 173–194.

Trépanier, M. & Chapleau, R. 2001 Analyse orientée-objet et totalement désagrégée des données d'enquêtes ménages origine-destination. *Revue canadienne de génie civil* **28** (1), 48–58.

Zhang, W., Miller, C. T. & DiGiano, F. A. 2004 Bacterial regrowth model for water distribution systems incorporating alternating split-operator solution technique. *J. Environ. Engng.* **130** (9), 932–941.