



This is a repository copy of *Case-based reasoning to support decision making for managing drinking water quality events in distribution systems*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/86843/>

Article:

Mounce, S.R., mounce, R.B. and Boxall, J.B. (2015) Case-based reasoning to support decision making for managing drinking water quality events in distribution systems. *Urban Water Journal*. Published online 21 May 2015. ISSN 1744-9006

<https://doi.org/10.1080/1573062X.2015.1036082>

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Case-based reasoning to support decision making for managing drinking water quality events in distribution systems

S. R. Mounce¹, R. B. Mounce² and J.B. Boxall³.

¹Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, Sheffield, UK, +44 (0)114 222 5750, S.R.Mounce@sheffield.ac.uk

(corresponding author)

² Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, Sheffield, UK, +44 (0)114 222 5750, richardmounce@hotmail.com¹

³ Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, Sheffield, UK, +44 (0) 114 222 5760, j.b.boxall@sheffield.ac.uk

Words: 5443 (plus 380 appendix) plus 7 figs (plus 5 in appendix) and 3 tables

This work was supported by the EPSRC under Grant EP/G029946/1.

¹ Current affiliation: Centre for Transport Research, MacRobert Building, King's College, University of Aberdeen, Aberdeen, UK, r.mounce@abdn.ac.uk

Case-based reasoning to support decision making for managing drinking water quality events in distribution systems

ABSTRACT

In order to better leverage past experience of water quality incidents, and to tap into the unique incident database currently being maintained and required by regulatory authorities, a data mining approach is herein proposed. The quality of drinking water is paramount to protecting public health. However water quality failures do occur, with some of the hardest to understand and manage occurring within distribution systems. In the UK, a regulatory process is applied in which water service providers must report on significant water quality incidents, their causes, actions and outcomes. These reports form a valuable resource that can be explored for improved understanding, to help with future incident management and evaluate potential solutions. Case-based reasoning is a knowledge-based problem-solving technique that relies on the reuse of past experience. The WaterQualityCBR software system presented here was developed as such a decision support tool to more effectively manage water quality in distribution systems.

KEYWORDS: Water quality; urban water management; decision support systems; data mining; drinking water

1. INTRODUCTION

The quality of drinking water is paramount to protecting public health and wellbeing. However water quality failures do occur, with some of the hardest to understand and manage occurring within distribution systems. The quality of drinking water in England and Wales has been regulated and standards enforced by the Drinking Water Inspectorate (DWI) since 1990. The DWI is a section of Department for Environment, Food and Rural Affairs (DEFRA), formed to guarantee regulatory obligations regarding drinking water quality to be followed by water companies. Through this position DWI

has the legal right to carry out technical audits of public water supplies (Gray, 2008). Davison et al. (2004) highlight the possible occurrence of a hazard throughout the water supply system, from catchment to consumers' taps. In drinking water distribution systems, when such a hazard occurs, it is considered as an incident. The term 'incident' is broadly defined by Hunter et al. (2003) as any circumstance in which there is a reasonable suspicion about the safety of water being supplied for drinking. More specifically, this term is described in the DWI annual report for 2002 as "an event affecting or threatening to affect drinking water quality" (DWI, 2002, p.137). The Water Industry (Suppliers' Information) Direction 2009 requires water companies to inform the Inspectorate of all events that have affected, or are likely to affect drinking water quality, or sufficiency of supplies and, where as a result, there may be a risk to consumers' health. According to Gray (2008), incidents can be classified as occurring in three parts of the water supply chain namely treatment works, service reservoir or in the distribution system. Incidents are generally classified into three different categories reflecting the fundamentals of water quality namely; physical (aesthetic), chemical and microbiological quality failure. In more recent years the proportion of incidents occurring within the distribution system has increased and includes microbiological contamination as well as discoloured water incidents (Gray 2008). The DWI assesses the water companies' response to the most significant incidents on an annual basis at a national level in England and Wales (provided in report form). The evaluation of water quality is based on information reported from the water companies. Water companies' responses and/or resolutions sometimes do not meet the expectation of the inspectorate and they can be fined if their performance is below a specified standard. The DWI also inspects the reliability of the test reports to ensure the quality of the water supplied (Clarke, 2003).

Modern water treatment works are extremely well managed, operated and monitored. Consequently, direct incidents originating there are rare. After leaving the treatment works, water is potentially subject to various mechanisms (Bates, 2000), such as microbial contamination (Tian et al., 2010) and to a variety of complicated reactions and interactions (Besner et al., 2001). Treated water experiences various changes whilst travelling through a distribution system and work has shown a link between the age of distributed water, persistent sporadic bacteriological failures and poor water quality in general (Machell et al. 2009, Machell and Boxall 2012). Distribution system infrastructure is an extremely complicated mix of materials, pipe sizes and structures connected in a network, usually in loops, developed in a piecemeal manner over long periods which acts as a dynamic physical, microbiological and chemical reactor, with a high surface area and with highly variable residence times. Consequently once water is in such an environment, increased risk of failures is much more likely than at the treatment works, and to understand how, when and why these occur is extremely complicated.

In summary, the distribution system is the last line of defence for water quality and a likely location for serious incidents. However, understanding and managing these complex, uncertain and ageing infrastructure systems is extremely challenging. Although failures can also occur at consumer's taps, these are not generally considered as part of the water supply infrastructure and their condition and operating regime are outside of water companies' control.

CBR has been proposed as a useful tool for providing diagnosis and solutions and hence it could potentially be used for the interpretation of, and guidance on, water quality incidents in water distribution systems. The aim of this research was to develop and demonstrate a decision support tool to illustrate how water companies could deal more effectively with water quality incidents (such as water discolouration or contamination) by using information from previous incidents. The tool manipulates a database (compiled in XML) of past significant events from several years DWI reporting. Such data mining of knowledge repositories is important in reducing the risk in reliance on so-called institutional memory and a limited (and reducing) number of domain experts.

2. CASE BASED REASONING

First developed in the late 1970s, case-based reasoning (CBR) is a knowledge-based problem-solving (memory based) technique that relies on the reuse of past experience (Kolodner, 1992). Examples are used directly to solve new problems, known as analogical reasoning in Artificial Intelligence. CBR is based on the assumption that similar problems from the past have similar solutions (since situations tend to repeat) and hence that solutions to new problems can be effectively addressed by reusing (and adapting) and learning from past solutions. According to Kolodner (1993), a case consists of three main parts: the problem, the solution and the outcome. The first element illustrates the current state of the world when the case was occurring. The second part states or derives the solution to the problem. The last part, which might be absent in some particular cases, shows the result after applying the solution. The CBR process comprises several steps: cases are stored in a database (the case-base) in a structured way (usually as a list of key attributes and associated values), then a

reasoning mechanism performs matching between these cases and a current situation or set of conditions and extracts from the case base the most similar instantiation. This case base should then provide useful information for the derivation of solutions in decision making. Adaptation and hybrid solutions can also be facilitated by the system and should be considered in decision making when evaluating the solution's suitability for the current problem. Additionally poor actions or responses to be avoided in decision making can be identified.

The stages of reasoning in CBR systems, based on cases, are known as classical R4-cycle (the four R's) as shown in Figure 1. This cycle consists of the following four main steps (Watson, 1999):

- **Retrieve** the case(s) from the case-base whose problem is most similar to the new problem.
- **Reuse** the solution(s) from the retrieved problem cases to create a proposed set of options for solution for the new problem.
- **Revise** the proposed solution to take account of the differences between the new problem and the problems in the retrieved cases.
- **Retain** the new problem and its revised solution as a new case for the case-base for future use if appropriate.

{Figure 1 approximately here}

CBR systems are particularly useful for diagnosis classification (identifying faults) and for prediction i.e. what happened when we saw this pattern before? They have been

applied in many domains and as of 2005, there were more than one hundred CBR systems mentioned in the literature (De Mantaras, 2005). Successful applications are to be found in spam filtering (Delany, 2005), the health sector (Xu 1994) and in law (Ashley 1992). Western legal systems depend on precedence and consulting case histories so the latter is a particularly appropriate field for CBR. Several examples are also evident in the literature for the water domain. CBR has been applied for water stress (Kukuric et al. 2008). Water-related cases consisted of indicators for natural conditions, stress(es) and mitigation measures. Policastro et al. (2004) applied a hybrid CBR system for monitoring water quality based on chemical parameters and algae population for evaluating samples taken from sites on different European rivers. Fenner et al. (2007) proposed a CBR methodology for modelling sewer infrastructure performance and condition. A nearest-neighbour retrieval algorithm for defining the similarity between cases was used and indices were used to assess outcomes. The methodology was tested on sewer data drawn from two UK water company regions.

3. APPLICATION OF CBR TO WATER QUALITY INCIDENTS

The CBR decision support tool that is reported here was developed to make use of the DWI annual reports for the period 2009-2011. These reports contain tabular summaries for each incident reported by a water company in an annex, albeit much of the text for outcomes is unstructured. The case-base was assembled in XML using an appropriate schema allowing further programmatic manipulation.

Expert subjective scoring of each DWI finding for each incident was conducted, as detailed later. A software tool called waterQualityCBR was developed as an MS Windows application in Visual Studio to perform the required aspects of CBR

functionality. The software has an intuitive GUI interface allowing for use by non-expert water company users. Figure 2 provides a schematic overview of the system operations, described further in subsequent sections.

{Figure 2 approximately here}

3.1. Construction of the case library

The properties of each incident are structured as attributes of a “case” and stored in the case-base. The selection of these attributes was based on all available data in the tabular summaries (Table 1 shows an example incident). The attribute fields (and types which are a mixture of numeric data and text strings) are described in the appendix.

The case-base required some manual resource to assemble, initially into a comma delimited file, then programmatically to a more flexible XML format. Note that some pre-processing/error checking was conducted which included standardisation of incident and cause type name. The case-base comprised 337 records of water quality incidents assembled in XML. Place and company names are ‘redacted’ or anonymised herein, to avoid commercial and/or public sensitivities, but this information is ultimately a useful part of the knowledge tool.

{Table 1 approximately here}

3.2. Evaluate and score DWI outcomes/ findings

The DWI findings are given as a number (1 to n in theory, though typically 2-4) of bullet point natural language comments. Some mechanism is necessary to infer management information from similar cases using such comments. This was achieved

by scoring the findings by using expert judgement (the authors for the case study). Each of the DWI bullet point natural language findings that appeared in the database cases were scored on a scale of 0 (very poor) to 10 (very good) with 5 being neutral. Note that findings were scored on an individual basis and not within the context of a particular case. Overall, since these were in response to incidents that were negative (i.e. had to be reported) it might be expected that there would be a bias to lower scores, as was found (the 614 outcomes for the case study dataset had a mean score of 3.5).

The overall outcome score for each case was then calculated as the average of the outcome scores of the DWI bullet point natural language findings for that case. This process contributed to an overall robustness that would be lacking in a per case score, by amalgamating scores across multiple outcomes per case (many outcomes repeated across cases). Effectively this process creates a regulatory assessment index for each case.

3.3. Case retrieval

The main objective of the CBR approach is to retrieve similar cases and present the user with a range of actions and interventions (good and bad) that have been used for similar incidents in the past such that their current decision making can be better informed. Similarity measures are generally used to achieve this by using any of the attribute fields from section 3.1 in the case-base library. A three-step approach was implemented as described in sections 3.3.1 to 3.3.3.

3.3.1. Search case-base for similar cases

There are two mechanisms provided in the software:

- a) Filter. Only cases that exactly match the criteria are returned. For example: population affected was greater than 10000; or that the case occurred at a certain time (months) of the year.
- b) Query. Cases that are above the match score threshold (described below) are returned (with a minimum and maximum on the number of cases returned). For example, specify a minimum of 5 (returning matches with a lower match percentage if necessary) and a maximum of twenty cases returned above a 95% match threshold.

For each variable i (search query terms), define the match score value $m_i(x_i, y_i)$ between query x and case y on a suitable scale, e.g. on a scale of zero to one or as a percentage. The overall match score is given by equation 1:

$$\frac{\sum_i w_i m_i(x_i, y_i)}{\sum_i w_i} \quad (1)$$

where w_i is the weight given to variable i . For numeric data, the match score is given by equation 2:

$$m_i(x_i, y_i) = \frac{|x_i - y_i|}{\max_i} \quad (2)$$

where \max_i is the maximum value for variable i , whereas for string data the match score is:

$$m_i(x_i, y_i) = \begin{cases} 1 & \text{if the query search term } x_i \text{ appears in the text } y_i \\ 0 & \text{otherwise} \end{cases}$$

For this application, with a limited number of string possibilities for the variables being searched, this approach is appropriate and provided an effective and efficient solution. String matching similarity metrics is a large area of research and techniques such as the

GATE (General Architecture for Text Engineering) suite of tools created by the University of Sheffield Natural Language Processing group (www.gate.ac.uk) could be considered if more sophisticated searching is required. For a large database the Boyer-Moore (Boyer and Moore, 1977) string search algorithm would provide a faster speed of search.

In order to perform the search, a query form was developed for maximum flexibility, and must first be completed as described in Appendix 1.

3.3.2. Rank actions and causes

Given all actions listed in the cases that are listed in the matches, the software calculates the average outcome score for each action (over those cases where the action was applied), e.g. if the action “flushed mains” was performed in two cases with outcome scores 5 and 7 respectively, then the action would be assigned an average outcome score of 6. In other words, this value shows the value across cases of applying an action (but only those returned by the search thus reflecting the subjectivity of the user’s settings for minimum/maximum cases returned at a particular threshold setting). These actions are ranked and displayed to the user. The number of occurrences of that action in the list of matching cases is also displayed. Also of interest to the decision maker is the most likely causes for a particular search / incident type. The user can perform a revised search if necessary, for example assuming a given cause type. Finally, they will select a number of actions to perform.

4. EVENT MANAGEMENT

4.1. Example

Consider a hypothetical situation of a microbiological contamination occurring at a treatment works, which is under the remit of XXXXX Water. The estimated number of affected residents is more than 10000. The query is completed as shown in figure A1.1. The results from the case-base are as shown in figure A1.2, with 15 cases matched at above 70% threshold. Note that the first three cases very closely match this new incident with match score approximately 97-99%. However, the three incidents are scored poor (3), medium (5.5) and high (7). Hence, the user may wish to examine the actions taken in the higher scored previous case (see the case window in figure A1.3). Similarly, a closely matched previous case with a poor score can be inspected to see what lessons can be learned from an inadequate response. Next, the rank actions function is performed and the results inspected to ascertain those actions that are performed regularly and/or with high score (see figure A1.4). Expert judgement will then dictate the particular response in this case, though shutting down the treatment works, sampling and any relevant repairs would likely feature in the course of action. The causes of this type of incident can also be reviewed, though for this situation the cause was unknown for all cases.

4.2. Validation

Knowledge validation is problematic in CBR and other knowledge based systems due to the nature of the task. Normally, CBR systems are allowed to learn by themselves as users enter a case, compare it with those in the case-base (retrieval) and when satisfied adds the case to the database. In this case, there is uncertainty as to what would have occurred had actions not been applied. Similarly, the recommended actions are being based on the DWI original findings.

It is possible to perform a type of validation by removing particular incidents from the database and not used in the algorithm for matching, then reviewing the results of searching the CBR system with respect to a removed incident case. Two similar discolouration incidents are present in the database: both affect approximately similar numbers of people and occur in the winter (see figure 3a and 3b), usefully one is poorly and the other highly rated in the DWI outcomes.

{Figure 3a approximately here}

{Figure 3b approximately here}

We perform two searches (with both these cases removed from the database):

- a) incident type = discolouration (weight 4), using 'Time of year' search and date = 28/1 (weight 1).
- b) Incident type = discolouration (weight 4), using 'Time of year' search and date = 29/12 (weight 1).

The top occurring actions (with above average score) for each case were sampled affected area, flushed mains, repaired mains and provision of bottled water (see figure 4 for the results for search a). We can see that in the case B recommended actions (i.e. three of the top four by occurrence) were followed (DWI score 8) whereas they were not in case A (DWI score 3). In other words, recommended actions for two similar (and previously unseen) events resulted in a positive DWI response when largely followed.

{Figure 4 approximately here}

5. STRATEGIC LEVEL RESULTS

Incident management in most water utilities is a well rehearsed process (Bradshaw et al. 2011) which is developed over time and hence there is a need for strategic level information that can provide steering on updating and optimising corporate processes. By utilising the full information in the case-base, updated annually with information from the DWI as described earlier, the system can provide summary information at a strategic level, for example to help inform policy or water company guidance documents.

The WaterQualityCBR software can be applied with appropriate filters to examine overall statistics. Figure 5 provides the breakdown of the causes of all the incidents in the case-base. Note that the cause for 48% of incidents is unknown (i.e. listed as unknown in the report to the DWI) and a burst main is the largest category of cause of serious water quality incident accounting for 15% of cases. The class ‘other’ contains all causes occurring less than twice. The second highest known single source of failures is ‘planned works’. This highlights the level of risk and uncertainty that currently accrue in water distribution systems and the need for better data in regard to both quality and quantity, as well as knowledge based systems that utilise this data.

Figure 6 provides the frequency for the top 20 most regular actions performed across the dataset, with the corresponding average rating index. The software can be used to drill down to optimum actions to perform in specific circumstances or in response to particular incidents. For example, if we change the query to interrogate only ‘discolouration’ incidents, the five most frequent actions in order of occurrence are:

sampled affected area (average index 4.3), flushed mains (4.4), repaired main (5.6), rezoned area (5.1) and provided bottled water (4.8). The selection of actions will of course require context specific evaluation of the incident as well as consideration of both occurrence and average rating index.

{Figure 5 approximately here}

{Figure 6 approximately here}

A more focussed query comparison on population affected can be useful to see what is effective at different scale. Two queries were run, one with target population 10 and one with 100000. We can see from table 2, which types of action were more likely to be performed (perhaps unsurprisingly actions like issuing boil water notices and providing bottled water are more common for smaller incidents) and also how the actions were viewed at that scale by the regulatory body (average rating index). Note that flushing mains gets the highest average rating index irrespective of scale. Viewing the corresponding causes from these searches, the two most likely causes (after ‘unknown’) for a small population affected were burst mains and cross connections with rainwater harvesting systems, whilst for large populations affected these were plant failure and planned work.

{Table 2 approximately here}

The tool can also be used to investigate seasonal phenomena and, by looking at the root causes of incidents, obtain a better understanding of them. Two searches were conducted on discolouration events: one in summer (target date 15/7) and winter (15/1). Parameters were selected to create clusters around these dates with weight 4 on the incident term 'discolouration' and threshold of 95%. Table 3 provides the results of these searches with the 'causes' function (top three causes).

{Table 3 approximately here}

There are a similar numbers of bursts causing (reported) discolouration in both seasons however there are certainly more bursts in the winter in the UK (Cook et al. 2005). Therefore we can conclude that there are fewer bursts resulting in reported discolouration during the winter. Hence, since the number of discolouration incidents is approximately equal it would appear that the water quality process is not driven by the winter freeze/thaw process that is thought to be responsible for a greater number of bursts in the winter. Hence, from a public health point of view this phenomena is not an issue and other processes such as corrosion or enhanced microbial activity during the warmer summer months are likely to be resulting in serious discolouration incidents.

Other possibilities for obtaining strategic guidance using the tool could include company specific data mining allowing overall performance between companies to be compared or against the average for particular causes and/ or incidents. The potential to explore and learn from this form of case base is clearly evident, together with the potential to tailor queries to yield knowledge at a variety of levels.

6. DISCUSSION

The DWI outcomes were scored using expert judgement (the authors). This procedure has implications for the significance to the outputs of the CBR. It could be questioned how easily (or otherwise) it was to judge on a 10 point scale based on DWI natural language comments. However, this scoring was performed based on how positively or negatively the DWI interpreted a water companies response and what impressed them, so was relatively robust albeit with an element of subjectivity. In practice these could be scored by an industry expert; although new scores could also be automated by matching to existing outcome-score combinations. It is important to note that each bullet DWI outcome comment was scored, not the overall incident. Thus with similar comments appearing in different cases, a level of consistency was achieved.

The CBR decision support tool makes use of cases (for all water companies) to recommend actions that can be taken to mitigate the effect of incidents as they arise. The tool utilises those incidents that must be reported legally to the DWI only, because this was the data publically available. Although water companies generally have guidelines for dealing with such incidents, it seems likely that the adoption of the CBR system could help improve performance. If sufficiently good records were kept at company level (of less significant events), then there would be scope to produce company specific versions of the software. The potential of this type of system could drive the establishment and maintenance of case bases both within the UK and internationally. In the UK, there is a consultation by data.gov.uk to make available the DWI water quality technical data and associated information as a sequel database (data.gov.uk 2014) which would facilitate data mining approaches such as this. The industry is moving towards risk based approaches, assessment of strategic issues (such

as rates of deterioration, failure causes and impact assessments) and needs new integrated operator tools for monitoring, mining, mitigation and management.

There are potentially risks with a successful system. It could create too strong a reliance and discourage thinking from first principles. The system could remove the perceived need for a human expert resulting in the loss of long term expertise. However, this type of CBR system is intended for decision support and not to supplant the expert human reasoning involved in responding to the specifics of an incident. Any given incident will involve a myriad of facts and circumstances that have to be taken into account, and can never hope to be captured in the CBR system. What the system does offer is a level of consistency in preliminary high pressure analysis and appraisal, such that more positive decisions should be made irrespective of operator specific experience and understanding. It is also important that the system does not negate the need for post event analysis and reporting, and that it is this more measured and considered analysis that actually provides the basis for long term learning within the CBR. The implementation of the Revise and Retain stages become imperative from this. If this software had sufficient take-up by water utilities these Revise and Retain steps of the CBR cycle could be incorporated in a formalised manner. The adaptation would be stored, and then based on the (yearly) findings and recommendations of DWI, the new case would be modified appropriately, and outcome scores Revised if necessary, and then stored in the case-base for future use once all information is available (Retain). Finally, the system can be used to enable enhanced strategic policy level analysis further reinforcing the need for human expertise, creative detailed analysis and interpretation from first principles.

7. CONCLUSIONS

This paper has described how a Case Based Reasoning tool has been developed as a decision support tool for water companies to potentially deal more effectively with water quality incidents by using information from previous incidents. The WaterQualityCBR software presented here manipulates an XML case-base of past significant events from several years DWI reporting. Benefits of the system include:

- Does not require knowledge of the problem domain and can provide solutions when no algorithmic method is available
- Can be employed for data mining and knowledge acquisition from the full case base by utilising the retrieval mechanism with general filters
- Can be utilised for operational event management providing solutions from similar cases from the past in the case-base and, importantly, ranking past actions in response to similar incidents
- Can provide information at a strategic level, for example to help inform water company policy or event response guidelines.

This CBR methodology can support decision-making and provide guidance for water utilities in managing drinking water incidents, which will ultimately reduce the risk to public health.

ACKNOWLEDGEMENTS

This work was supported by the Pipe Dreams project funded by the UK Science and Engineering Research Council, grant EP/G029946/1. The authors would particularly like to thank Mr. Hieue T. Nguyen for data preparation and initial experimentation. They also acknowledge the DWI in making publically available the annual reports.

REFERENCES

Ashley, K. D., 1992. Case-Based Reasoning and its Implications for Legal Expert Systems. *Artificial Intelligence and Law*, 1, 113-208.

Bates, A., 2000. Water as consumed and its impact on the consumer--do we understand the variables? *Food and chemical toxicology*, 38 (1), 29-36.

Besner, M.C., Gauthier, V., Barbeau, B., Millette, R., Chapleau, R. and Prevost, M., 2001. Understanding distribution system water quality. *Journal American Water Works Association*, 93 (7), 101-114.

Boyer, R. S. and Moore, J. S., 1977. A Fast String Searching Algorithm. *Comm. ACM* (New York, NY, USA: Association for Computing Machinery), 20 (10), 762-772.

Bradshaw, R., Gormley, A. M., Charrois, J. W., Hruday, S.E., Cromar, N. J., Jalba, D. and Pollard, S. J. T., 2011. Managing incidents in the water utility sector – towards high reliability? *Water Science and technology: water supply*, 11 (5), 631-641.

Clarke, K., F., 2003. The British experience. In: Agthe, D. E., Billings, R. B. and Buras, N. (eds) *Managing urban water supply*. Netherlands: Kluwer Academic Publishers, 241-257.

Cook, D. M., Boxall, J. B., Hall, S. J. and Styan, E., 2005. Structural integrity and water quality in water distribution networks. *Proceedings of CCWI 2005*, University of Exeter.

Data.gov.uk, 2014. Drinking Water Inspectorate (DWI) sequel database system [online]. Available from: <http://data.gov.uk/dataset/drinking-water-inspectorate-dwi-sequel-database-system> [Accessed 10/1/2014].

Davison, A., Howard, G., Stevens, M., Callan, P., Kirby, R., Deere, D. and Bartram, J., 2004. Water safety plans. World Health Organisation, Geneva.

Delany, S. J., Cunningham, P. and Coyle, L., 2005. An Assessment of Case-Based Reasoning for Spam Filtering. *Artificial Intelligence Review Journal*, 24 (3-4), 359-378, Springer.

De Mantaras, R.L., Mcsherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M.L., Cox, M.T., Forbus, K., Keane, M., Aamodt, A. and Watson, I., 2005. Retrieval, reuse, revision and retention in case-based reasoning. *Knowledge Engineering Review*, 20 (3), 215-240.

DWI., 2002. Annex3: Definitions and glossary terms [Online]. Available: <http://dwi.defra.gov.uk/about/annual-report/2002/annex3.pdf> [Accessed 28 July 2012].

Evans, W.A., 1994. Approaches to intelligent information retrieval. *Information processing and management*, 7 (2), 147–168.

Fenner, R. A., McFarland, G. and Thorne, O., 2007. Case-based reasoning approach for managing sewerage assets. *Proceedings of the Institution of Civil Engineers Water Management*, 160, 15-24.

Gray, J., 2008. Water contamination events in UK drinking-water supply systems. *Journal of water and health*. 6 (1), 21-26.

Hunter, P., Andersson, Y., Von Bonsdorff, C., Chalmers, R., Cifuentes, E., Deere, D., Endo, T., Kadar, M., Krogh, T. and Newport, L., 2003. Surveillance and investigation of contamination incidents and waterborne outbreaks. *Assessing microbial safety of drinking water*. Edited by A. Dufour, M. Snozzi, W. Koster, J. Bartram, E. Ronchi and L. Fewtrell. ISBN 92 4 154630 1 (WHO) and 1 84339 036 1 (IWA Publishing) © WHO OECD 2003.

Kolodner, J.L., 1992. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6 (1), 3-34.

Kolodner, J., 1993. *Case-based reasoning*. Morgan Kaufmann.

Kukuric, 2008. *The i3s document series: Using 'case based reasoning' for the solution of water stress problems* [online].

Available from: <http://i3s.aquastress.net/tools/CBR/Aquastress%20I3S-Case%20Based%20Reasoning%20Users%20Guide.pdf>. [Accessed 20 May 2012]

Machell, J., Boxall, J. B., Saul, A. J., Bramley, D., 2009. Improved Representation of Water Age in Distribution Networks to Inform Water Quality. *Water Resources Planning and Management ASCE*, 135 (5), 382-391.

Machell, J. and Boxall, J. B., 2012. Field Studies and Modeling Exploring Mean and Maximum Water Age Association to Water Quality in a Drinking Water Distribution Network. *Water Resources Planning and Management ASCE*, 138 (6), 624–638.

Policastro, C. A., Carvalho, A. C. P. L. F. and Delbem, A. C. B., 2004. A Hybrid Case Based Reasoning Approach for Monitoring Water Quality. *Innovations in Applied Artificial Intelligence, Lecture Notes in Computer Science*, 3029, 492-501.

Tian, Y., Liu, Y. and Liu, H., 2010. Measurement and analysis on biofilm bacteria of the pipe wall in water distribution system. *IEEE 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE), China*, 1-4.

Watson, I., 1999. Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems*, 12 (5), 303-308.

Xu, L.D., 1994. Case based reasoning. *Potentials, IEEE*. 13 (5), 10-13.

Appendix 1. Software details

The attribute fields in the XML database (and types which are a mixture of numeric data and text strings) are as follows:

- water company (string) – the water company name (utility company in England and Wales)
- incident type (string) – the type of incident
- cause type (string) – the suspected cause of incident, if known
- population affected (integer) – total population at risk from the incident
- date (datetime) – of event
- time span (string) – duration in days or hours
- area (string) - location
- actions (list of strings) – actions undertaken by water company in response to incident
- outcomes (list of strings) – DWI findings on this incident

In order to perform the search, a query form was developed for maximum flexibility, and must first be completed as illustrated in Figure A1.1. Ideally searches should be driven by multiple filters and queries. Some query options allow logarithmic scaling to be used for numeric data, such as for population. The bottom three parameters allow fine tuning of the number of matches returned. In particular, the threshold percentage gives the cut off on how closely cases must match (from equation 1). However, a minimum number of matches can be specified such that a set number will be returned below this threshold percentage if necessary.

{Figure A1.1 approximately here}

Once the search executes, a list of matches is produced as shown in Figure A1.2. These matching cases will then be utilised for recommending actions. The case window displays additional information about the case (shown in Figure A1.3).

{Figure A1.2 approximately here}

{Figure A1.3 approximately here}

Given all actions listed in the cases that are listed in the matches, the software calculates the average outcome score for each action. These actions are ranked and displayed to the user. The number of occurrences of that action in the list of matching cases is also displayed – see figure A1.4. Simultaneously presenting both the average score for the action (when applied in this instance) and the number of occurrences allows the user to evaluate actions from both perspectives. The action list can be sorted on either score or occurrences.

{Figure A1.4 approximately here}

Also of interest to the decision maker is the most likely causes for a particular search / incident type. Figure A1.5 shows this screen.

{Figure A1.5 approximately here}

Figure captions

Figure 1: CBR R4 cycle adapted for application to water quality incidents

Figure 2: WaterQualityCBR software system main operations
Figure 3a: Discolouration incident A

Figure 3b: Discolouration incident B

Figure 4: Action scoring and occurrence screen (search a)

Figure 5: Causes of reported water quality incidents

Figure 6: Most popular actions from case-base along with average rating index

Figure A1.1: The query form

Figure A1.2: List of matches

Figure A1.3: The case window

Figure A1.4: Action scoring and occurrence screen

Figure A1.5: Causes screen

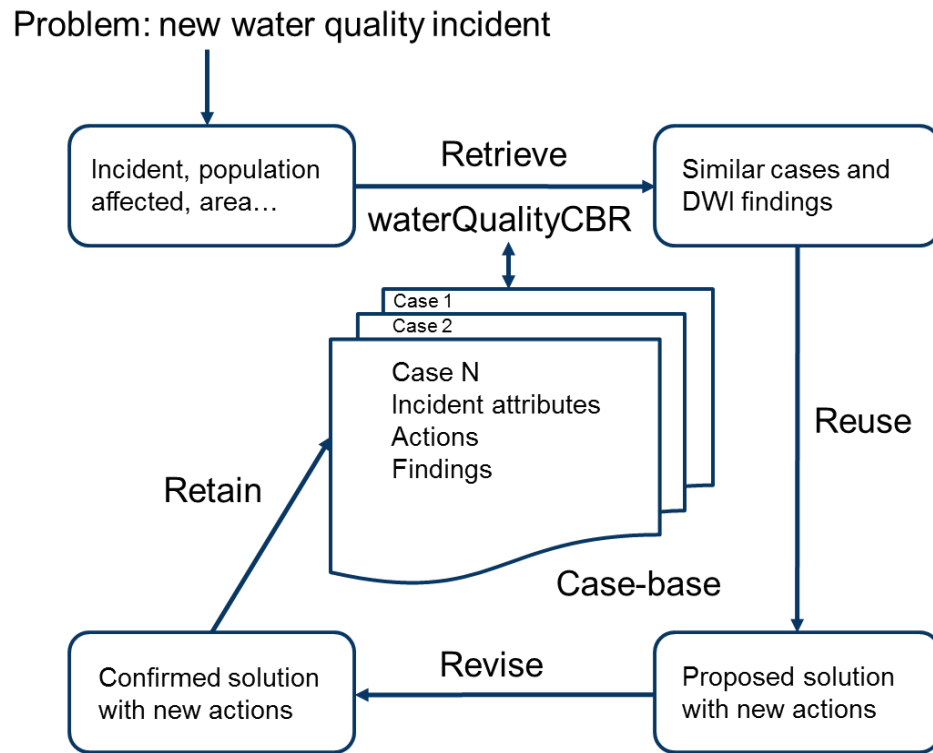


Figure 1. CBR R4 cycle adapted for application to water quality incidents

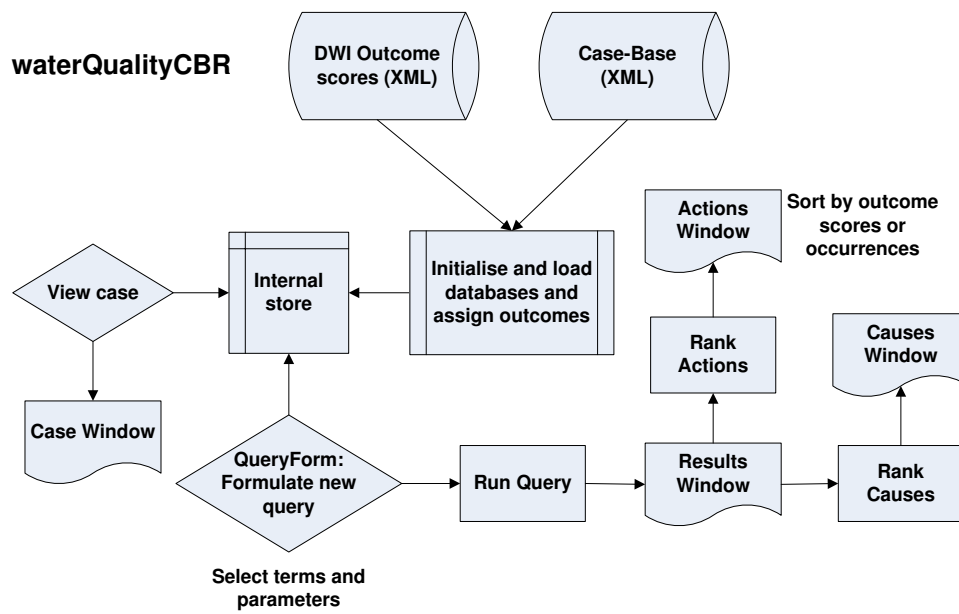


Figure 2. WaterQualityCBR software system main operations

The screenshot shows a software window titled "CaseWindow" with a table of incident details. The table has two columns: a label column and a value column. The values for some fields are redacted with black boxes. A "Back" button is located in the bottom right corner of the window.

Water company	[REDACTED]
Incident type	brown discolouration
Cause type	burst main
Population affected	12500
Date	28/01/2009
Duration	12:00:00
Area	[REDACTED]
Actions	Repaired main
Outcomes	Inadequate procedures
	Inadequate contingency planning
Outcome score	3

Figure 3a. Discolouration incident A

CaseWindow	
Water company	[REDACTED]
Incident type	brown discolouration
Cause type	burst main
Population affected	16900
Date	29/12/2010
Duration	1.00:00:00
Area	[REDACTED]
Actions	Supplied bottled water
	Flushed mains
	Rezoned area (brought in water from different source)
	Collected samples
Outcomes	Satisfactory
	Results of further investigative sampling were satisfactory and showed that company actions were effective in returning supplies to normal in a timely ma...
Outcome score	8

Back

Figure 3b. Discolouration incident B

Action	Score	Occurrences
Updated geographical information systems (GIS) records	10.00	1
The company is planning to replace the service reservoir soo...	10.00	1
Repaired main	5.88	8
Flushed mains	5.28	9
Provided bottled water on request	5.04	4
Sampled affected area	4.62	19
Rezoned area (brought in water from different source).	4.00	2
Shut down treatment works	4.00	1
Provided an alternative supply by tanker/bowser	4.00	1
Rezoned area (brought in water from different source)	3.67	3
Risk assessment undertaken but not documented	3.50	1
Review of procedures	3.50	3
Flushed and repaired mains	3.50	1
Retrained staff	3.50	1
Acted promptly and effectively to resolve the issue	3.00	1
Advised consumers to trickle flush their taps	3.00	1
Briefed key customers on need to communicate any water s...	3.00	1
Undertook a risk assessment in advance of the planned work	3.00	1
Introduced further treatment	2.75	1
Reviewed drinking water safety plan	2.75	1
Optimised treatment	2.75	1
Carried out water fittings inspections	2.50	1
Rezoned area (brought in water from different area)	2.50	1
Planning to replace a common copper supply pipe serving a ...	2.50	1

Back

Figure 4. Action scoring and occurrence screen (search a)

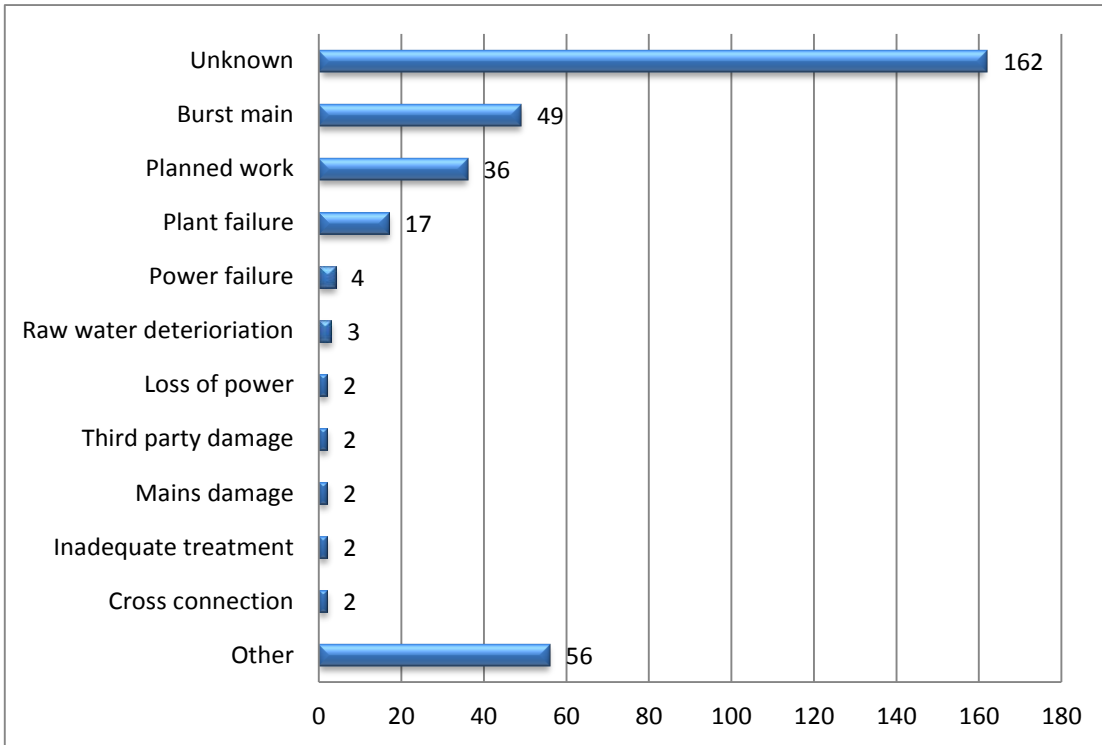


Figure 5. Causes of reported water quality incidents

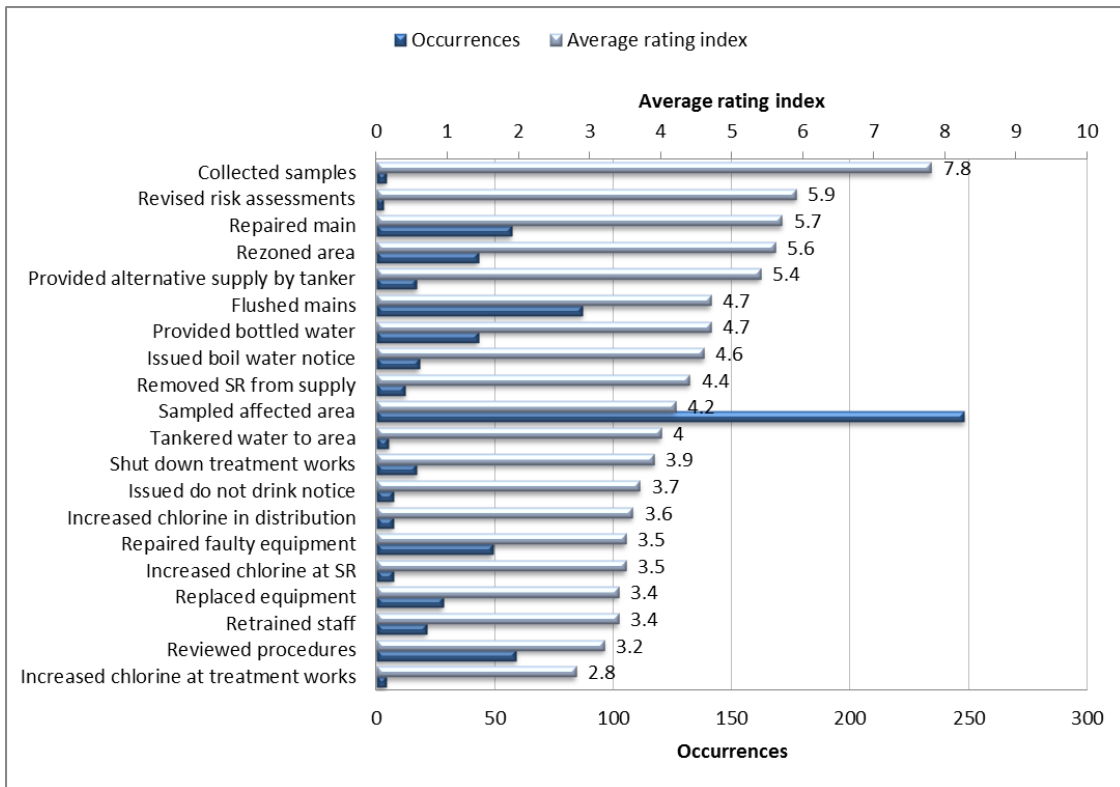


Figure 6. Most popular actions from case-base along with average rating index

The screenshot shows a window titled "QueryForm" with a standard Windows-style title bar. The main area contains a table-like structure for building a query. The columns are: Variable, Filter value, Filter option, Query value, Query option, and Query weight. The rows are: Water company, Incident type, Cause type, Population affected, Date, Duration, Areas, Actions, Outcomes, and Outcome score. Each row has input fields and dropdown menus. For example, "Water company" has a filter value field, a dropdown set to "or", a query value field with a blacked-out area, a dropdown set to "or", and a query weight field. "Incident type" has a query value of "microbiological". "Population affected" has a query value of "10000" and a dropdown set to "normal". "Date" has a dropdown set to "time of year". "Duration" has a dropdown set to "normal". "Areas" has a query value of "works" and a dropdown set to "or". "Actions" has a dropdown set to "and". "Outcomes" has a dropdown set to "and". "Outcome score" has a query weight field. Below the table, there are three rows of settings: "Minimum number of matches" with a value of 5, "Maximum number of matches" with a value of 20, and "Threshold percentage" with a value of 70. To the right of these settings are two buttons: "Run" and "Back".

Variable	Filter value	Filter option	Query value	Query option	Query weight
Water company		or		or	
Incident type		or	microbiological	or	
Cause type		or		or	
Population affected			10000	normal	
Date				time of year	
Duration				normal	
Areas		or	works	or	
Actions		or		and	
Outcomes		or		and	
Outcome score					

Minimum number of matches: 5
 Maximum number of matches: 20
 Threshold percentage: 70

Run Back

Figure A1.1. The query form

Water company	Incident type	Cause type	Population	Date	Duration	Area	Outcome score	Match score
	microbiological co...		28000	01/12/2009	1.00:00:00		5.50	99.89
	microbiological co...		229400	26/03/2010	1.00:00:00		3.00	98.64
	microbiological co...		360000	29/10/2010	1.00:00:00		7.00	97.82
	microbiological co...		7000	17/07/2010	1.00:00:00		3.00	74.98
	microbiological co...		470	06/08/2009	3.00:00:00		4.00	74.94
	microbiological co...		113	17/09/2010	2.00:00:00		4.33	74.94
	loss of disinfection		39165	10/05/2010	04:00:00		10.00	74.82
	microbiological co...		48154	17/06/2010	1.00:00:00		4.00	74.76
	cryptosporidium c...		50000	08/10/2009	3.00:00:00		4.00	74.75
	microbiological co...		70044	16/12/2010	3.00:00:00		4.00	74.63
	microbiological co...		74875	11/08/2010	2.00:00:00		0.00	74.60
	microbiological co...		88000	02/07/2010	1.00:00:00		3.00	74.51
	microbiological co...		250000	11/03/2010	1.00:00:00		4.00	73.51
	microbiological co...		250035	10/08/2010	1.00:00:00		2.67	73.51
	cryptosporidium		517000	16/01/2010	1.00:00:00		10.00	71.85

Figure A1.2. List of matches

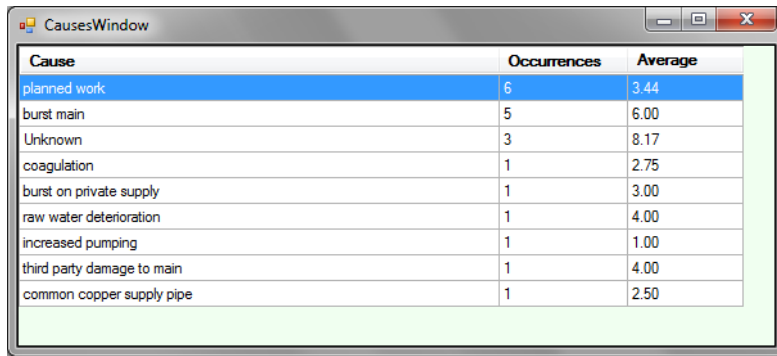
Water company	[REDACTED]
Incident type	microbiological contamination
Cause type	
Population affected	360000
Date	29/10/2010
Duration	1.00:00:00
Area	[REDACTED] works
Actions	Shut down treatment works Desludged the dissolved air floatation (DAF) plant Confirmed that there was no issue with sampling facilities
Outcomes	Satisfactory The company was asked to confirm that all the materials of construction and jointing of the sample line and tap comply with Regulation 31 of the Water ...
Outcome score	7
Match score	97.8233830845771

Back

Figure A1.3. The case window

Action	Score	Occurrences
Subsequent final water samples contained no <i>Cryptosporidium</i> oocysts	10.00	1
Repaired faulty pre-ozonation equipment	10.00	1
Retrained staff	10.00	1
Shut down treatment works	7.50	3
Repaired faulty equipment	7.00	2
Confirmed that there was no issue with sampling facilities	7.00	1
Desludged the dissolved air floatation (DAF) plant	7.00	1
Thorough treatment investigation carried out and no issues identified	5.50	1
Carried out enhanced monitoring on return to service	5.50	1
Removed service reservoir from supply	4.33	1
Provided bottled water on request	4.00	1
Flushed mains	4.00	1
Issued a boil water notice	4.00	1
Inspected final water reservoirs	4.00	1
Repaired main	4.00	1
Repairs made to on-site blending tanks and pipework	4.00	1
Inspected the reservoir internally and externally and cleaned the res...	4.00	1
By ♦ passed service reservoir	4.00	1
Increased chlorine residual leaving the service reservoir	4.00	1
Company is currently introducing flow alarms for ferric dosing which ...	4.00	1
Reviewed analytical procedure and analytical controls	4.00	1
Reviewed works performance	4.00	1
Sampled affected area	4.00	9
Collected samples	3.00	1
Found sample line had additional connections to equipment presenti...	3.00	1
Bypassed service reservoir	3.00	1
Investigated works	3.00	1
Replaced faulty equipment	3.00	1
Identified potential ingress to the contact tank	2.67	1
Reservoir inspection	0.00	1

Figure A1.4. Action scoring and occurrence screen



The screenshot shows a window titled "CausesWindow" with a table containing the following data:

Cause	Occurrences	Average
planned work	6	3.44
burst main	5	6.00
Unknown	3	8.17
coagulation	1	2.75
burst on private supply	1	3.00
raw water deterioration	1	4.00
increased pumping	1	1.00
third party damage to main	1	4.00
common copper supply pipe	1	2.50

Figure A1.5. Causes screen

Table captions

Table 1: Example water quality incident in raw form from DWI report

Table 2: Top actions for small versus large population affected (mean rating index 3.5)

Table 3: Top three causes for summer versus winter discolouration

Date and duration	Area	Estimate of population affected	Nature and cause of the event	Main actions and findings from the Inspectorate investigation
20 Jan 2009 For 36 hours	XXXXX	1000	Brown discolouration due to planned work.	<p>The Inspectorate classified this event as significant.</p> <p>Water Company action:</p> <ul style="list-style-type: none"> • Sampled affected area. • Review of procedures. • Repaired main. <p>DWI comments and findings:</p> <ul style="list-style-type: none"> • Inadequate risk assessment. • The Inspectorate was critical that no flushing was undertaken.

Table 1. Example water quality incident in raw form from DWI report

Query on population affected 10			Query on population affected 100,000		
Action	No	Average Rating Index	Action	No	Average Rating Index
Sampled affected area	26	4.3	Sampled affected area	23	3.6
Flushed mains	13	4.8	Repaired faulty equipment	8	2.9
Provided bottled water	8	3.6	Review of procedures	7	3.1
Issued a boil water notice	7	4.2	Flushed mains	5	4.8
Issued a do not drink notice	6	3.8	Replaced faulty equipment	4	2.6

Table 2. Top actions for small versus large population affected (mean rating index 3.5)

Summer			Winter		
Cause	No	Average Rating Index	Cause	No	Average Rating Index
Burst main	9	6.1	Burst main	8	6.4
Planned work	4	2.5	Planned work	6	3.4
Unknown	3	5.8	Unknown	4	7.3

Table 3. Top three causes for summer versus winter discolouration