Application of CBR for Intelligent Process Control of a WWTP

Josep Pascual-Pañach[a,b,c,1], Miquel Àngel Cugueró-Escofet[a,b], Miquel Sànchez-Marrè[c], Pere Aguiló-Martos[b]
[a] *CCB Serveis Mediambientals, SAU*
[b] *Consorci Besòs Tordera*
[c] *Knowledge Engineering and Machine Learning Group (KEMLG)*
Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC)
Universitat Politècnica de Catalunya · BarcelonaTech

Abstract. This paper proposes the use of a Case-Based Reasoning (CBR) system for the control and the supervision of a real wastewater treatment plant (WWTP). A WWTP is a critical system which aims to ensure the quality of the water discharged to the receiving bodies, stablished by applicable regulations. At the current stage the proposed methodology has been tested off-line on a real system for the control of the aeration process in the biological treatment of a WWTP within the ambit of *Consorci Besòs Tordera* (CBT), a local water administration in the area of Barcelona. For this purpose, data mining methods are considered to extract the available knowledge from historical data to find a useful case base to be able to generate set-points for the local controllers in the WWTP. The results presented in this work are evaluated taking into account the performance of the CBR method e.g. case base size, CBR cycle time or number of cases resolved satisfactorily (forthcoming steps will include on-line tests). For this purpose, some Key Performance Indicators (KPI) are designed together with the plant manager and process experts, in order to monitor key parameters of the WWTP which are representative of the performance of the control and supervision system. Hence, these KPI are related with water quality regulations —e.g. ammonia concentration in the WWTP effluent— and the economic cost efficiency —e.g. electrical consumption of the installation. In order to evaluate the results, different flat-based memory organizations (i.e. cases are stored sequentially in a list) for the case base are considered. First, a unique case base is used. At the current stage and for the results shown in this work, this case base is divided in multiple libraries depending on a case classification. Finally, the combination of this approach with Rule-Based Reasoning (RBR) methods is proposed for the next stages of the work.

Keywords. Case-Based Reasoning, Intelligent Process Control, Wastewater Treatment Plant, Data Mining.

1. Introduction and context of the work

Many disciplines take advantage of the use of models of real-world processes in order to get useful insight of the corresponding real systems' behaviour, e.g. Environmental

[1] Josep Pascual Pañach, R&D Department, Consorci Besòs Tordera, 241 Sant Julià Avenue, 08103 Granollers (Barcelona), Spain; E-mail: jpascual@besos-tordera.cat.

Decision Support Systems (EDSS). Former EDSSs used *mechanistic models*, but increasingly available huge amounts of data gathered from these systems triggered the use of new empirical models. *Empirical models* are based on direct observation, measurement and extensive data records. The first empirical models used were mathematical and statistical methods e.g. Multiple Linear Regression (MLR) models. The success of several inductive *machine learning techniques* within the Artificial Intelligence (AI) area led to their application in EDSSs. Some instances of this usage are e.g. the Association Rules (AR) models, Classification Rules (CR) models, Decision Tree (DT) models or Bayesian Network (BN) models. Since the 80s, both the former mathematical/statistical empirical models and the later machine learning empirical models have been named *data mining methods*, because they result from a mining process using these data. With the use of *data mining models* within the AI framework, the EDSS have evolved to Intelligent Environmental Decision Support Systems (IEDSSs) [1]. IEDSSs integrate knowledge stored by human experts through years of experience in a certain environmental process operation and management. In addition, data can be mined through the intelligent analysis of available large databases coming from historical operation of this environmental process. Thus, *knowledge integration and data mining model production*, as well as *reasoning* and *interoperation* among the models produced, are key steps and a challenge to build reliable IEDSSs. Moreover, for each different type of Environmental system, the corresponding IEDSS is developed in an *ad-hoc* basis, with no general framework available for the automation of their deployment [2].

On the one hand, *Interoperability* is defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [3]. Regarding this issue, XML (eXtensible Markup Language) is a meta-language that provides one of the most effective ways to interchange information between several software components and share the corresponding information semantics. Furthermore, in the field of data mining, the Data Mining Group [4] is an independent, vendor led consortium that develops data mining standards, such as the Predictive Model Markup (PMML) Language. PMML is a standard for statistical and data mining models, supported by over 20 vendors and organizations. PMML uses XML to represent data mining models, supporting the most common ones, e.g. AssociationModel, RegressionModel, TreeModel, RuleSetModel, NeuralNetwork, ClusteringModel. In addition, some work related with model and data integration and reuse in EDSS may be found in [5], and an overview of model integration is presented in [6]. An interesting work in this area is also presented in [7], where the Drools Rule-based integration platform is used as a unified data model and execution environment, and in [8], where a general framework for the development of interoperable IEDSSs was proposed.

On the other hand, workflows are graphical notations, which were first introduced to model and describe Business Processes [9]. The use of *Visual workflows* can be a very helpful tool for specifying the workflow involving all the steps from the raw data to the end of the process defined, including the models produced, the model executors and other auxiliary processes. The idea of using workflows for the control of a process e.g. in organisations, has been pointed in the literature [10].

Until now, in most cases the interoperability of the models is achieved by ad-hoc interactions, which may be considerably improved. Despite there are some architecture proposals in the literature to solve the interoperation of different models, there is no

common framework to implement Interoperable IEDSS, which would allow an easy integration and (re)use of different AI or statistical/numerical models in a whole IEDSS.

2. Control and Supervision Approach

This works proposes an alternative to avoid ad-hoc approaches for the design of control and supervision schemes of sanitation systems, particularly WWTPs. The design of the control and supervision tool depends not only on the available processes in the WWTP –e.g. organic matter removal, phosphorus removal–, but also on the plant design –e.g. treatment capacity, plant type, available sensors and actuators. Hence, huge amounts of time and resources are invested in different stages of the development, from design, implementation and start-up to maintenance.

The approach presented here aims to reduce the implementation time of the WWTP control and supervision strategy. Traditionally, these are based on non-graphical programming environments, which are not especially suited for model-based design and which are challenging to implement and to maintain, especially for certain applications like the one considered here. Alternatively, an IDSS approach based on data-models using a visual programming approach is proposed and applied. The concept of the proposed architecture and IDSS is described in [2]. Here, we focus on the case-based reasoning algorithms proposed to produce the corresponding outputs —i.e. set-points for the control of the different processes involved— and its implementation based on a real WWTP facility.

2.1. System architecture

The main goal of the IPCS presented here is to generate the set-points for the local controllers and the decision support system. The architecture of the whole system is shown in Figure 1. The tool developed here is based on a three-layer architecture presented in [2]. The WWTP is controlled and supervised using the process workflow layer shown in Figure 2, which is application core, by means of different data-driven models. Considering the nature of the real application, the need for complete validated and reconstructed datasets is paramount in order to apply further methods using these data. This situation is especially challenging in sanitation systems, where sensors and corresponding communication systems are exposed to rough conditions —e.g. operation in extreme environments with presence of dirtying agents— which seriously jeopardize their operation and may produce higher rates of malfunction, leading to higher amounts of potentially non-reliable raw data, e.g. outliers or missing values. Hence, a data validation and reconciliation stage is paramount to prevent this behavior, e.g. the one in [11] or further data mining methods to produce valid data for models generation. These models can be e.g. rule models induced from decision trees or case databases. All models interoperate in the process control workflow to supervise the system by discriminating between abnormal situations and normal operation and to control the process by generating actuator set-points based on knowledge obtained from data. Rule models can also include rule patterns which may complement the case database with valid new cases, when the CBR model is incomplete or distorted by a high number of potentially invalid cases —not identified and removed from the case database—, which are often challenging to be filtered when working with real

measured raw data. It may also include human expert knowledge of the system, so it is worth to consider an easy user interface to integrate such human-based knowledge.

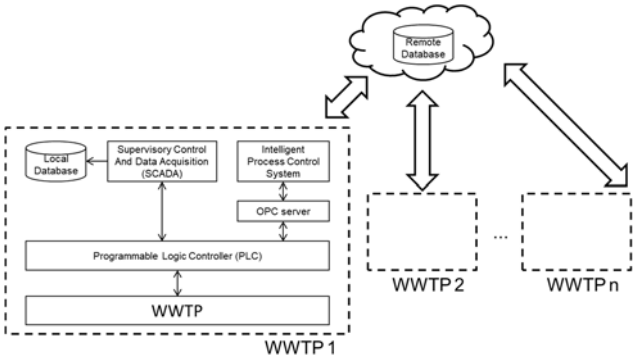

Figure 1 System architecture

The designed tool implements the classical CBR scheme shown in Figure 3. The CBR cycle is implemented using Simulink to comply with the specifications described in section 3, the use of visual workflows and programming languages. The result is a set of tools that consist on different methods that can be used in different CBR cycle stages to create the most appropriate application for each sanitation system.

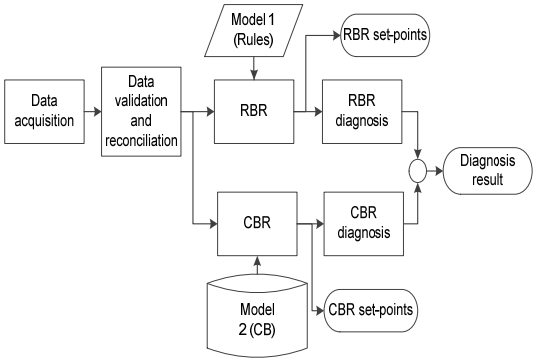

Figure 2 Process Control workflow

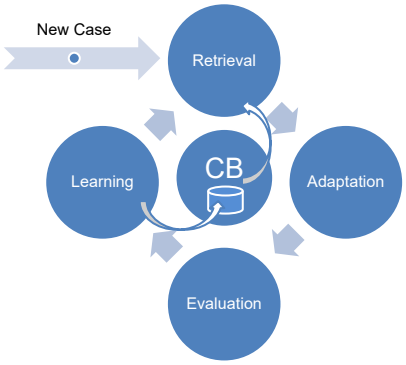

Figure 3 Case-Based Reasoning cycle scheme

3. Case study: Granollers WWTP

3.1. Description

Consorci Besòs Tordera (CBT) is a local water administration composed of 64 municipalities in four different regions of Catalonia with a population of about 470.000 inhabitants. CBT is the responsible for the sanitation facilities from the very beginning in project and building stages to the final facilities operation and maintenance — including 300 km of sewers and 23 WWTPs— with the main objective of preserving and improving the good health of the rivers in its area. All WWTPs within CBT ambit

are based on the activated sludge process for the wastewater treatment. Plants capacity range from 1000 m^3/day to over 40000 m^3/day, or expressed in other terms, from some hundreds of population equivalent (PE) to over 300000 PE. Each WWTP includes water and sludge lines, and in some cases, a biogas line. Even though this similar layout, there are some particularities that imply a custom-made control system, e.g. number and type of actuators and sensors or the influent characteristics.

The aeration of the biological reactor is one of the most important processes to be supervised, since is the most critical and resource consuming for the whole WWTP. The aim of this process is to remove organic matter (organic carbon) and nutrients (nitrogen and phosphorus) from the sewage water. To make this biological process feasible oxygen is necessary. Oxygen is usually introduced from the environment to the biological reactors by using aeration blowers. The process of removing nutrients and organic matter is a two-step process that requires periods with oxygen and periods without oxygen, so the adequate management of the aeration blowers is very important. In a first stage, nitrogen is oxidized to nitrates in presence of oxygen (nitrification). This process is done by some autotroph bacteria, which consume dissolved oxygen. Then, in a second stage, nitrates are reduced to gaseous state nitrogen (denitrification) by some kind of heterotroph bacteria. In an anoxic situation, these bacteria use nitrates instead of dissolved oxygen and consume carbon obtained from organic matter. Other controlled processes are e.g. phosphorus removal by chemical dosing or bypass flow in case of overload situations.

Hence, the first pilot considered for the approach presented here has been focused on a real WWTP in the Barcelona area (i.e. Granollers WWTP), in order to generate the set-points for the local controllers of this process. In this particular case, there are two biological reactors working in parallel and sharing the same air distribution system.

3.2. Available data and preprocessing

As introduced in Section 2, a basal problem for CBR designing —or any other data-based method—is the quality of the input data. Generally, process data is stored in the local database of the control and supervision system of each WWTP. The available data gathered by installed sensors for the case of study considered is listed in Table 1, while Table 2 shows the outputs generated by the system presented —i.e. set-points for local control loops. All these sensors are measuring online data stored with one minute sample time.

Table 1 Available sensors used to generate the case base

Type of sensor	# sensors	Units
Dissolved oxygen	10	mg/l
Plant Input flow	1	m^3/h
Biological reactor input flow	1	m^3/h
Internal recirculation frequency converter	2	Hz
External recirculation flow	2	m^3/h
Ammonia concentration	2	mg/l
Nitrate concentration	2	mg/l
Suspended Solids concentration in the biological reactor	2	mg/l
Suspended Solids concentration in the output	1	mg/l
Phosphorus concentration in the output	1	mg/l
Aeration valves position	10	%
Air pressure	1	mbar

Table 2 CBR system outputs

Solution	# solutions	Units
Oxygen set-points	10	mg/l
Valves position set-points	10	%
Pressure set-point	1	mbar
Biological reactor state (nitrification or denitrification)	2	-

The first step is to retrieve historical data from the database. With the main objective of considering information representing all the possible process behaviours, including for example seasonal behaviours —e.g. temperature changes have important impact on the biological process performance—, a recent dataset gathered from the year 2018 is considered. The first problem detected is the huge amount of data when considering one minute sample time —the number of values gathered by each sensor in a one year period are about 500,000—, which could be avoided since generally time constants of the processes considered are remarkably higher than one minute. Hence, at the current stage of the CBR system design, a sample time of ten minutes is considered. Thus, for one year period the number of values gathered per sensor is reduced by a rate of 90 % of the original dataset, i.e. 50,000 values per sensor, without losing relevant information. Second, as commented previously, the quality of these data is often low due to different causes, namely missing values, inconsistency between values with the same time stamp, out of range values or abnormal behaviours caused by faulty sensors or maintenance operations. These behaviours suggest the use of a data validation and reconciliation stage in order to provide reliable and complete datasets. After detecting problems in data there are two options to deal with them: the most radical one is to delete those cases that are not valid because of its quality; the second one is trying to replace wrong values with possible valid values using auxiliary information from, for example, other sensors or other techniques like the ones described in [11]. In a first version of the case base to be tested with the real process the most radical option is used: all cases with low quality are deleted and only when the inconsistency problems can be solved with the use of other reliable signals, wrong values are corrected. For example, pressure values can be crossed with blower state (switched on or switched off) and electrical consumption signals; or valves position can be crossed with the corresponding measured flow.

After the data pre-processing process described above, the production of a valid set of data is assumed. Then, considering the knowledge of the plant manager and experts on the process other auxiliary features are calculated, e.g. trends of some of the sensors described in Table 1 or moving averages. Finally, a case base with 58 features in the descriptive part and about 19,000 cases is obtained (the 65% of the cases available before the data pre-processing have been deleted).

3.3. Case-based reasoning system design

The case base obtained from historical data after the pre-processing steps is firstly organized in a flat memory, in a unique case base. Later, in an improved version is organized as a multiple case bases depending on the biological reactors state in order to reduce the execution time and to be more accurate in the retrieval process. The biological reactor state —nitrification or denitrification—does not always depend on the plant situation. The main feature to take into account to determine in which phase has to be operated the process is the ammonia concentration. However, sometimes

depending on the specific situation and the criterion of the plant manager, the nitrification or denitrification steps can be forced. For the plant described in this work —with two biological reactors— and considering two states of the biological process —i.e. nitrification and denitrification, the use of four case bases is proposed. In the retrieval step 1 case is retrieved from the case memory, according to the used similarity measure. As all the variables of the dataset are numeric, the Euclidean distance is used as a first approach, but other approaches will be considered.

In the adaptation stage the same solution of the most similar case is applied, i.e. null adaptation. The range of all set-points is defined by the experts on the process and can be changed by the plant manager. If the given solution is not within the range, set-points are adapted to fulfill the configured limits. In order to be more accurate in the set-point calculation, the use of rule-based adaptation methods taking into account the plant manager knowledge will be explored in the next steps.

Regarding the four stages of the CBR (Figure 3), the evaluation of the CBR output and the learning methods implementation are in progress, so results shown in this work are related to retrieval and adaptation stages. Both the evaluation of the CBR output and the learning stages are paramount for the forthcoming online tests, since the quality of the output provided to the actual plant and the method to keep updated and validated information for future actuation is key in order to assure current and future quality of operation. In Section 3.4, some KPIs are discussed to evaluate the performance of the CBR system. In Section 4, some ideas about how to deal with these stages are given.

3.4. Results

The prototype of the proposed CBR system is tested in an off-line fashion using real data gathered from the WWTP Hence, here the prototype is running in the plant and receiving new cases to obtain a solution using data from the real process, but the proposed solutions are not used in the WWTP operation, i.e. they are only compared with the ones generated by the current on-line control system. Figure 4 shows some oxygen concentration set-points for Reactor 1 and a 48 hours scenario, whereas the complete test length is two weeks. Figure 5 shows some valves position set-point for the same reactor and period. Figure 6 shows the Graphical User Interface (GUI) used to test the prototype of the designed IDSS in the WWTP of the case study.

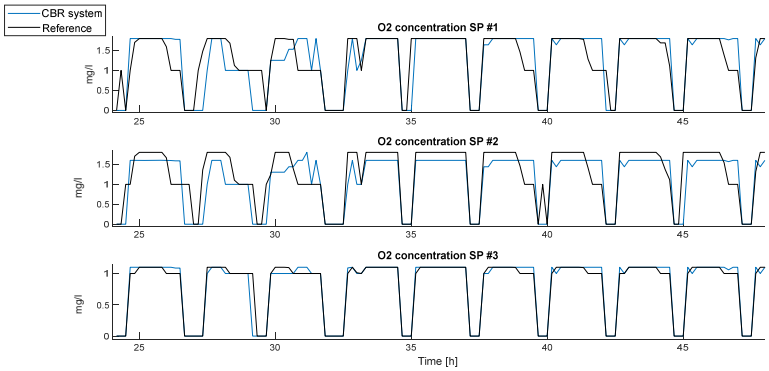

Figure 4 Oxygen set-points generated for Reactor 1 vs. reference

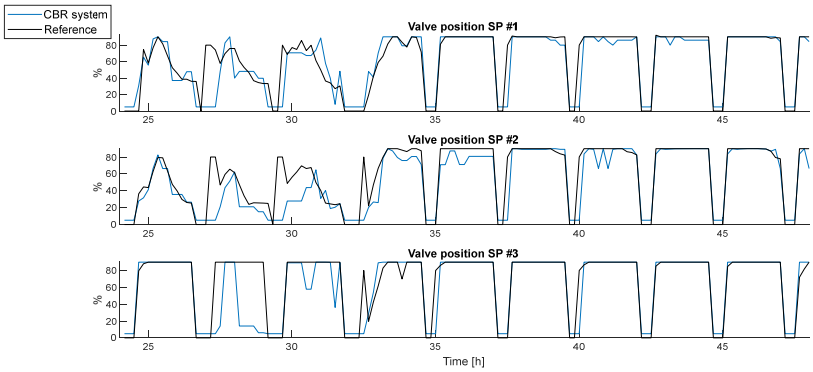

Figure 5 Valves set-points generated for Reactor 1 vs. reference

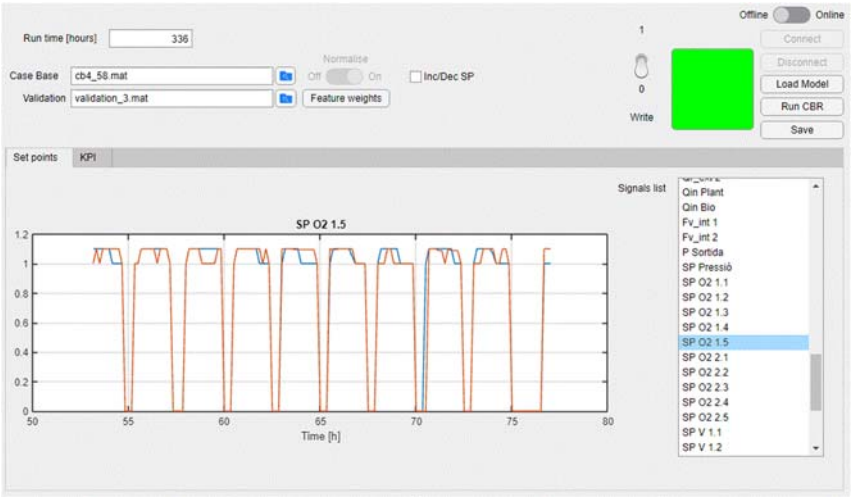

Figure 6 GUI of the online application developed using Matlab

As it can be seen the performance of the CBR system in the offline test is promising for further on-line tests of the tool. In most of the cases the solution given is very similar to the reference one, although there are some wrong predictions in some situations involving predicted set-point values that are quite different from the actual set-points values. The CBR performance is evaluated in terms of the accuracy of the generated set-points compared to the reference ones, considering a tolerance of the 10 % (i.e. for lower differences the CBR system set-point is considered successful, otherwise unsuccessful). Taking this into account, the global accuracy for the two weeks scenario considering the set-point generation of 21 actuators (Table 2) is 67 %. In the case of oxygen set-points, the mean accuracy is 62 %, while valves position set-points accuracy is higher, about 75 %. Incorrect case retrieval may be caused by e.g. failed stored cases in the case base or the lack of information to distinguish situations with the available features. On the other hand, it is possible that the dynamic nature of this domain implies that the problem cannot be solved with high accuracy using an individual case retrieval. The solution might depend not only in the current situation but also on how this situation has been reached, so *temporary dependences* with past

cases should be taken into account, as proposed in [12] with the use of a temporary CBR approach.

The execution time of the CBR cycle or the size of the case base are also important parameters, regarding the on-line nature of the target real application. The tool is running on a 64 bit Windows Server machine with three cores CPU and four GB RAM. At the current stage, the execution time of the CBR cycle (from the reading of all measurements to the calculation of the generated set-points) is about 0.26 seconds, ranging from 0.08 to 1 second. Considering that the process is executed every minute, no improvement of the execution is needed in order to provide the output on time. Regarding the case base, note that the size of the case base model is about 12 MB.

To the light of the current results, future steps include on-line test stages in the actual operation of the WWTP. In addition to the conclusions drawn from the offline test, it can be highlighted the importance of having mechanisms to evaluate the solution obtained from the point of view of the process performance, since it is known that a given set-point sometimes does not have the expected effect on the system. Solutions in the past should have the same effect in a similar situation at present. But again, it could be some differences between them that cannot be distinguished with the available information to describe the domain, or even that the sensors or actuators involved are not giving the appropriated measurements or responses. To this end, two levels are proposed for the evaluation phase: first, by defining some KPI to determine that the solution applied is having the desired effect on the system. The KPIs suggested here are related to the outflow quality and the WWTP efficiency. At the current stage, these KPI —which are designed together with the plant manager— are:

- 24 h moving average (MA) of ammonia concentration: This value is stablished by applicable regulations to a maximum of 4 mg/l.
- Blower electrical consumption: Historical daily average consumption is used as a threshold to be compared with the current daily average consumption.
- Nitrogen removal efficiency: Total nitrogen in the influent and in the effluent of the WWTP is not an online measure but an offline analytic measure obtained three times per week. Since nitrogen removal efficiency is stablished at 80 % by applicable regulations, it is also proposed as a useful performance indicator.

In the daily operation and considering the slow dynamics of the processes involved, these indicators can be useful in a medium-term horizon and may be involved with expert criterion —validating e.g. the daily set of solutions—, but further indices should be used to evaluate the solutions in a short term horizon. Thus, the second level of evaluation will include some rules related with the actuators response, e.g. consistency among set-point and opening position or expected valves air flow.

4. Conclusions and future work

In this paper the application of a CBR approach to solve the control of the aeration in the biological process of a real WWTP is presented. The concept architecture presented in [2] has been applied in a real case study using real measured operation data in an off-line fashion. The CBR off-line system has been tested using real data obtained from the actual process, but with no actuation in the actual facility, which will be performed in the following stage. Some KPIs are designed together with the plant manager and proposed to evaluate the CBR output in the forthcoming online tests. Obtained results

are promising to devise further steps towards this on-line implementation but still some problems have to be solved e.g. incorrect case retrieval that may occur due to failed stored cases or possible missing information in the case base or to the temporary and continuous nature of this domain.

Thus, further work includes the use of rule-based methods to improve the adaptation and evaluation phases and the definition of the learning strategy, as well as the inclusion of high-performance data validation and reconciliation stages and the Rule-Based Reasoning (RBR) methods in the scheme in order to provide redundancy and complement the CBR outputs. In addition, episode-based reasoning modules are been developed to take into account the temporary nature of the system.

Acknowledgements

The authors acknowledge the partial support of this work by the Industrial Doctorate Programme (2017-DI-006) and the Research Consolidated Groups/Centres Grant (2017 SGR 574) from the Catalan Agency of University and Research Grants Management (AGAUR), from Catalan Government.

References

[1] Sànchez-Marrè, M., Gibert, K., Sojda R., Steyer J.P., Struss, P. and Rodríguez-Roda, I. (2006) Uncertainty Management, Spatial and Temporal Reasoning and Validation of Intelligent Environmental Decision Support Systems. 3rd International Congress on Environmental Modelling and Software (iEMSs'2006). iEMSs' 2006 Proceedings, pp. 1352-1377.

[2] Pascual-Pañach, J., Cugueró-Escofet, M.À., Sànchez-Marrè, Aguiló-Martos, P., 2018. An Interoperable Workflow-based Framework for the Automation of Building Intelligent Process Control Systems, in: 9th International Congress on Environmental Modelling and Software "Modelling for Sustainable Food-Energy-Water Systems." Fort Collins, USA.

[3] Institute of Electrical and Electronics Engineers (1990) IEEE Standard Computer Dictionary: a Compilation of IEEE Standard Computer Glossaries. New York.

[4] DMG (2014) The Data Mining Group (http://www.dmg.org) leads the development of the Predictive Model Markup Language (PMML). Current version PMML 4.2. February 2014.

[5] Rizzoli, A. E., Davis, J.R. and Abel, D.J. (1998) Model and Data Integration and re-use in Environmental Decision Support Systems. Decision Support Systems 24:127-144, 1998.

[6] Argent, R.M. (2004) An Overview of Model Integration for Environmental Applications-components, frameworks and semantics. Environmental Modelling & Software 19:219-234, 2004.

[7] Sottara, D., Bragaglia, S., Mello, P., Pulcini, D., Luccarini, L and Giunchi, D. (2012) Ontologies, Rules, Workflow and Predictive Models: Knowledge Assets for an EDSS. 6th International Congress on Environmental Modelling and Software (iEMSs'2012). iEMSs' 2012 Proceedings, pp. 204-211.

[8] Sànchez-Marrè, M. (2014) Interoperable Intelligent Environmental Decision Support Systems: a Framework Proposal. 7th International Congress on Environmental Modelling and Software (iEMSs'2014). iEMSs' 2014 Proceedings, Vol. 1, pp. 501-508.

[9] ter Hofstede, A.H.M, van der Aalst, W.M.P., Adams, M., Russell N., (2010) Modern Business Process Automation. Springer.

[10] zur Muehlen, M. (2004) Workflow-based Process Controlling. Berlin: Logos-Verlag.

[11] Cugueró-Escofet, Miquel À. et al. 2016. A Methodology and a Software Tool for Sensor Data Validation/Reconstruction: Application to the Catalonia Regional Water Network. Control Engineering Practice 49: 159–72. http://linkinghub.elsevier.com/retrieve/pii/S0967066115300459 (September 29, 2016).

[12] Sánchez-Marré M., Cortés U., Martínez M., Comas J., Rodríguez-Roda I. (2005) An Approach for Temporal Case-Based Reasoning: Episode-Based Reasoning. In: Muñoz-Ávila H., Ricci F. (eds) Case-Based Reasoning Research and Development. ICCBR 2005. Lecture Notes in Computer Science, vol 3620. Springer, Berlin, Heidelberg