A Sensor and Semantic Data Warehouse for Integrated Water Resource Management

Andreas Abecker[1], Torsten Brauer[1], Babis Magoutas[2], Gregoris Mentzas[2], Nikos Papageorgiou[2], Michael Quenzer[1]

Abstract

The goal of the EU-FP7 project WatERP is to achieve more interoperability of software systems along the water-supply chain. A central element to achieve that is the WatERP Water Data Warehouse which shall act as a central data-exchange platform between different software systems. The WDW shall be able to store and provide sensor, measurement and forecasting data, as well as semantic knowledge about the water-supply chain. It shall be as standards-compliant as reasonable and shall offer query and reasoning facilities over sensor data, spatial information and ontological knowledge. In this paper, we show the basic architecture of the WDW and shortly discuss its main design decisions.

1. Motivation

A holistic view on the way of a drop of potable water, starting with a precipitation in a river catchment area or with a deep-well to an aquifer, and ending with a private household, an industrial or agricultural consumer, reveals a whole bunch of administrative, commercial and societal actors and stakeholders that altogether create the drinking-water supply chain or value-network, or have some interest in it. Depending on the respective regulatory, administrative and economic system in a country, the specific form of this value-network differs significantly from country to country. Nevertheless, it is always composed of a number of different actors (if we also consider private consumer households, even a huge number) with significantly differing ways of working, depending on their place and their role in the water-supply chain. Such actors can be, e.g., bulk-water suppliers, dam administrations, local or regional utility companies, large industrial consumers, private households, but also environmental agencies, NGOs, etc. Seen from the legal perspective, these actors comprise also different kinds of organizations, public authorities, administration unions, special-purpose associations, private companies, public-owned enterprises, public-law institutions, private persons, etc. Also the tasks related to the water-supply chain are not restricted to providing water to consumers, but also affect constraints and secondary goals, like energy production, flood protection, environment protection, and – of course – quality control and quality assessment of the delivered drinking water.

It is a nearby hypothesis that some of the activities in such a multitude of actors, actions and objectives could gain efficiency or effectiveness if there was a comprehensive and prompt exchange of data between different stages in the supply chain. For instance, when optimizing the water distribution in temporarily droughty areas, an early estimation of the upcoming water yield of a bulk-water supplier at a given interconnection point together with a good short-term/mid-term estimation of the different competing consumers' demand could facilitate the reduction of causes of conflict. Furthermore, economic incentives (like flexible water tariffs), organizational measures (timed coordination between different industrial consumers), motivational activities (giving information or appeals to the citizen) or regulations (setting of quotas) – as methods of *Demand*

[1] Disy Informationssysteme GmbH, 76131 Karlsruhe, Germany, firstname.lastname@disy.net

[2] National Technical University of Athens, 15780 Athens, Greece, [elbabmag|gmentzas|npapag]@mail.ntua.gr , Institute for Communication and Computer Systems (ICCS)

Management – could be better coordinated. Or, in the case of a foreseeable peak-demand, the water production can be controlled more efficiently (operation management of deep-wells, water works, reservoirs, etc.). But also in water-rich areas, optimization potentials can be expected, in particular with respect to the energy consumption and/or energy costs (in the case of flexible energy tariffs) needed for operating the water-distribution network (pumps for filling elevated tanks and for keeping the water pressure), if one takes into account more exact demand forecasts and/or more actual consumption data or if one coordinates the consumption-plans of large industrial customers.

However, such a systematic, comprehensive and real-time data and software interoperability throughout the whole water-supply chain is not given at all, at the time being. Typically, if there is communication/coordination between two actors, then it is between two immediately subsequent stages in the value-chain, it is done for the immediate usage in a work process, and it is often realized with simple, often tailor-made, technical means, and not always refers to most actual data.

In order to improve this situation, the EU-FP7 project **WatERP** ("Water-Enhanced Resource Management – Where Water Supply Meets Demand") [1, 2]

- on one hand, realizes forecasting and decision-support tools for optimizing the whole water-supply chain (tools for demand forecasting, for energy-optimized network operations, for demand management, etc.), and
- on the other hand, implements a data-exchange platform through which the different actors in the water-supply chain can share their observation data, forecasted values, and management decisions.

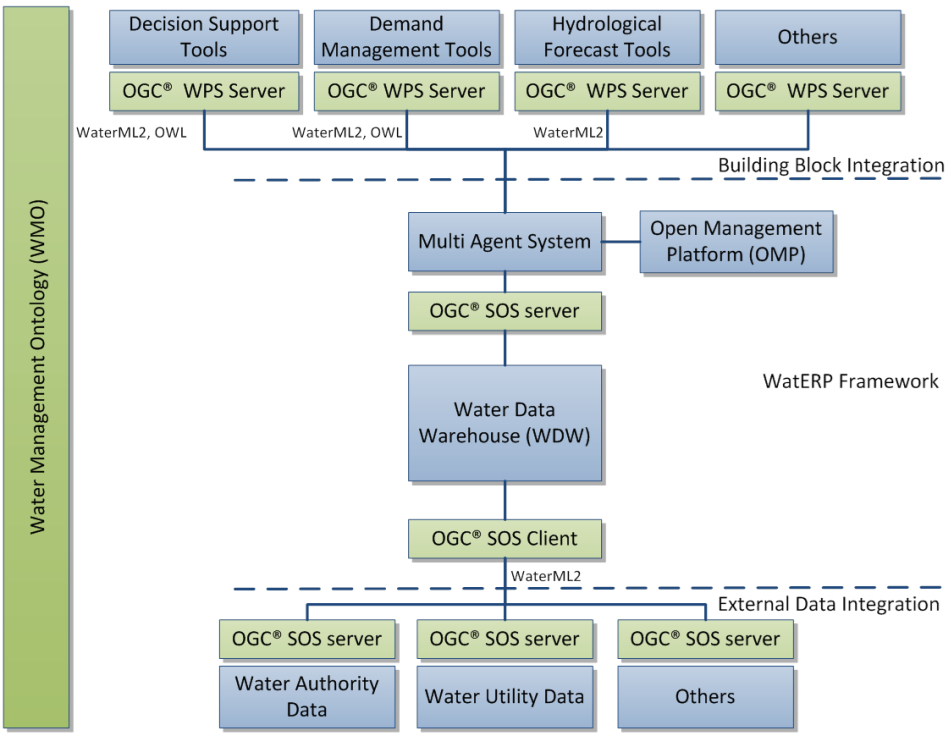

Figure 1: Overall WatERP Software Architecture

Here, it should also be noted that optimization potentials do not necessarily need to span *different* organisations; often enough, there are already interoperability deficiencies between the software systems of only *one* actor.

The main element of the WatERP data-exchange platform is the so-called **WatERP Water Data Warehouse (WDW)**. In the WDW, different kinds of data, information and knowledge created in

the water-supply chain is centrally collected and stored and is provided for different tools and purposes in different views, formats, aggregations, etc.

Figure 1 gives an overview of the overall WatERP software architecture: The WDW stands at the center of the system. It is fed by data-producing systems linked to the WDW through OGC SOS interfaces exchanging WaterML2.0 data. The data-consuming, decision-support tools realize a Service-oriented Architecture (SOA), enabled by the OGC WPS (Web Processing Service) protocol for communication, and facilitated by a Multi-Agent Systems which helps to find appropriate services and orchestrate complex processing workflows. The conceptual backbone of the system is the WatERP Water Management Ontology (WMO) which represents all decision-relevant aspects of the water-supply chain as an OWL knowledge base (see also [1, 2]).

This paper is structured as follows: In Section 2, we describe the overall architecture of the WatERP WDW and introduce some main design decisions. In Section 3, these design decisions are discussed with some more detail. Finally, we conclude with Section **Fehler! Verweisquelle konnte nicht gefunden werden.**.

2. Architecture of the WatERP Water Data Warehouse

The Water Data Warehouse copies data from operational systems and offers them in optimized views, aggregations and formats for analysis purposes. Typical functionalities of a Data Warehouse comprise data harmonization, measuring fault assessment, data cleansing, as well as purpose-specific selection, view and aggregation mechanisms. Figure 2 sketches the WDW architecture within the overall project context. One major design decision was to devise **two separate storage areas**:

- **Mass data** (time series describing sensor measurements about hydrology, network operation, meteorology, …) is typically highly repetitive, very simply structured, but coming in large volumes. Large amounts of this kind of data can easily and efficiently be managed with "conventional" database technologies, in our case, an *object-relational storage area* (*PostGIS* database). These data populate a data schema which is based on the structure of the OGC / ISO conceptual model „*Observations and Measurements*"[3] and on OGC *WaterML2.0*[4], respectively.
- On the other hand, in order to describe all decision-relevant aspects of a water supply chain comprehensively – which might be needed to allow interoperability of formerly isolated systems responsible for specific complex tasks – one might have a need to describe more irregularly structured knowledge, **complex relationships** between objects, definitions of technical terms or interrelationships of them, or generic relationships which abstract away from specific facts (like rule-based knowledge or arithmetic relationships). For expressing such more complex issues, computer science has developed knowledge-representation languages for expert systems which have been consolidated and standardized through "Semantic Web" technologies. In the Semantic Web area, so-called triple stores have been introduced for storing and processing complex knowledge. More recently, *triple stores* have been extended towards geospatial reasoning which allows drawing logical deductions that also include also simple notions of spatial relationships and spatial deductions. For instance, sensors could be geo-referenced and the system could, if appropriately modelled, find all sensors in a given administrative district or in the state which comprises this district; or all sensors situated on the rivers that a given public authority is responsible for. So, the second main storage area of the WatERP WDW is a triple store enabled to do geospatial reasoning (*OWLIM* triple store with

[3] http://www.opengeospatial.org/standards/om

[4] http://www.opengeospatial.org/standards/waterml

SesameAPI together with the *uSeekM* library to add geospatial search functions and geospatial reasoning through the *GeoSPARQL* query mechanism).

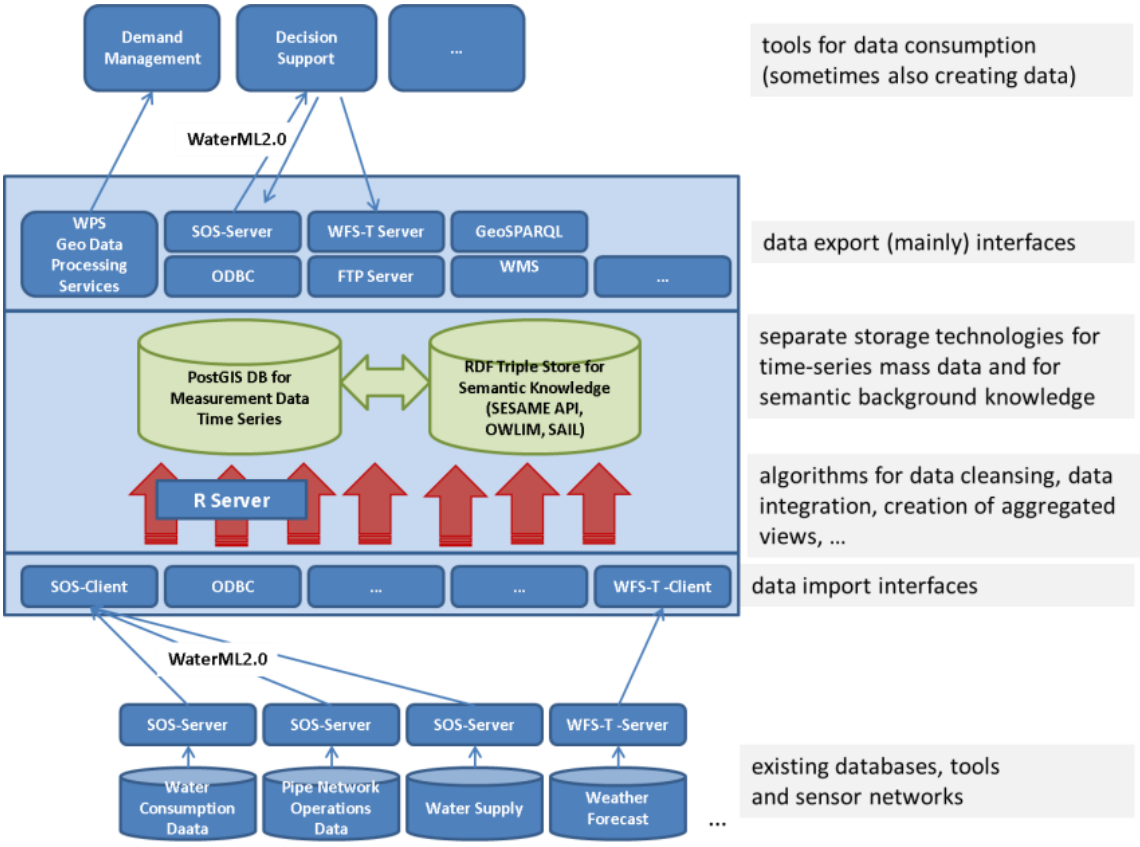

Figure 2: Simplified Architecture of WatERP Water Data Warehouse

A second important design decision was to base all developments to the most possible and reasonable extent on the **open standards of the Open Geospatial Consortium (OGC)** which is the most important data and software interoperability organization in the geodata area. Hence, all sensor and measurement transmission activities as well as the internal organization of the time-series PostGIS database follow the OGC WaterML2.0 specification.

Thirdly, it should be noted that many data pre-processing steps can excellently be realized through the **R[5] statistics software** which has also extensions to directly interoperate with SOS sensor data streams.

All these three design decisions are explained with some more detail in the following section.

3. Discussion of Specific WDW Design Aspects

3.1. WaterML2.0 and SOS

All sensor and measurement transmission activities follow the **OGC SOS protocol** with **WaterML2.0 as data schema**. The internal organization of the mass-data storage also implements a simplified and tailor-made data schema based on the OGC WaterML2.0 data specification. For implementation, we employed the 52°North SOS Server.[6] All data sources are linked to the WDW through SOS – except for the meteorological forecasts where we considered WFS/WFS-T more

[5] http://www.r-project.org/

[6] http://52north.org/communities/sensorweb/sos/

suited for representing the structure of the data. Of course, time-series data are also offered through SOS/WaterML2.0. In our practical experience, up to now, the existing infrastructure works stable and reliable. Regarding parallel processing of many sensors, the solution seems to work very well. We discovered one problem of the 52°North SOS Server in the situation that one would try to send as one large package a huge number of observations (> 1 Mio) with the full WaterML2.0 metadata. Of course, this is not the usual use case for SOS (where a data stream comes in package-wise step-by-step), but it could happen in the case of importing a larger legacy observation database. In such a case, the server would probably crash. But, of course, dividing the database into smaller packages or loading as a CSV file would help. In general, the OGC SOS protocol with the WaterML2.0 data format turned out to be a powerful and stable infrastructure in our experience.

3.2. Semantic Knowledge Base and Geospatial Reasoning

The incorporation of a semantic knowledge base follows the current trend to empower modern software solutions by knowledge-based components, to increase interoperability through ontologies and to provide data with *Semantic Web* methods according to the *Linked Open Data* paradigm. However, a complete "semantification" (representation, storage and processing of all aspects with Semantic Web methods and tools) of all data in WatERP seemed not feasible and promising to us, especially regarding the measurement data. Instead, the time-series data and the semantic knowledge complement each other and could together be exploited for powerful analyses and queries about the considered water supply system. For instance, the structured relationships in the triple store might be used to logically describe a water-distribution network and its geospatial aspects, as well as some background knowledge, for instance, about measurement methods for assessing water quality. Then, specific water-quality measurements could continuously be fed as time series into the PostGIS database. The metadata for measurements would have references into the knowledge base. This allows making complex queries which combine time-series sensor data and semantic background knowledge. For instance, one could ask for all measurements made in a certain geographic area and using an analytical sensor technology with certain characteristics; or, about all sensor data from a certain point on in the water supply chain which make statements about a certain group of related chemical or biological pollutants. Analysing the answers could help to find upstream dischargers responsible for a contamination and take appropriate counter-measures, or it could help to identify endangered spots further downstream and take suitable protection or purification measures.

It should be noted that, as a side-effect of the overall architecture, the simultaneous evaluation of quantitative, measurement-data based and qualitative, knowledge-based conditions in *one* query cannot be done *within* the WDW framework. So, if one would like to find, for example, all sensor observations coming from a certain administrative area (spatial reasoning) and made with a certain type of sensor technology (structural reasoning) which are above a certain threshold (sensor value), this would have to be implemented as a query agent outside the WDW that merges the query results of the SOS sensor-data query and the logic-based query.

Figure 3 shows the implementation of the WDW geospatial reasoning which is based on the uSeekM und Sesame Java libraries in combination with the PostGIS geodatabase and the Sesame-based RDF triple store OWLIM. **Sesame**[7] is a Java framework for the implementation of RDF stores that offers an extensible API on top of which other stores can be built. OpenSahara **uSeekM**[8] is a Java library which realizes spatial indices and spatial queries as an extension of Sesame Java based triple stores. uSeekM creates a separate R-Tree index for spatial data. It catches

[7] http://www.openrdf.org/

[8] http://www.w3.org/2001/sw/wiki/USeekM

GeoSPARQL[9] queries and rewrites them as a combination of queries against the spatial index and the RDF triple store. Different triple stores can be used, as far as they support Sesame Sail. The **Sesame Sail** API allows for functional extensions of RDF stores as a low-level system API for RDF stores and inference engines which abstracts from implementation details and so allows to use different stores and inference engines. In WatERP, we employ **OWLIM**[10] as an RDF store that allows reasoning over an OWL ontology, such that, altogether, the WDW Triple Store allows combined reasoning over ontological and spatial knowledge.

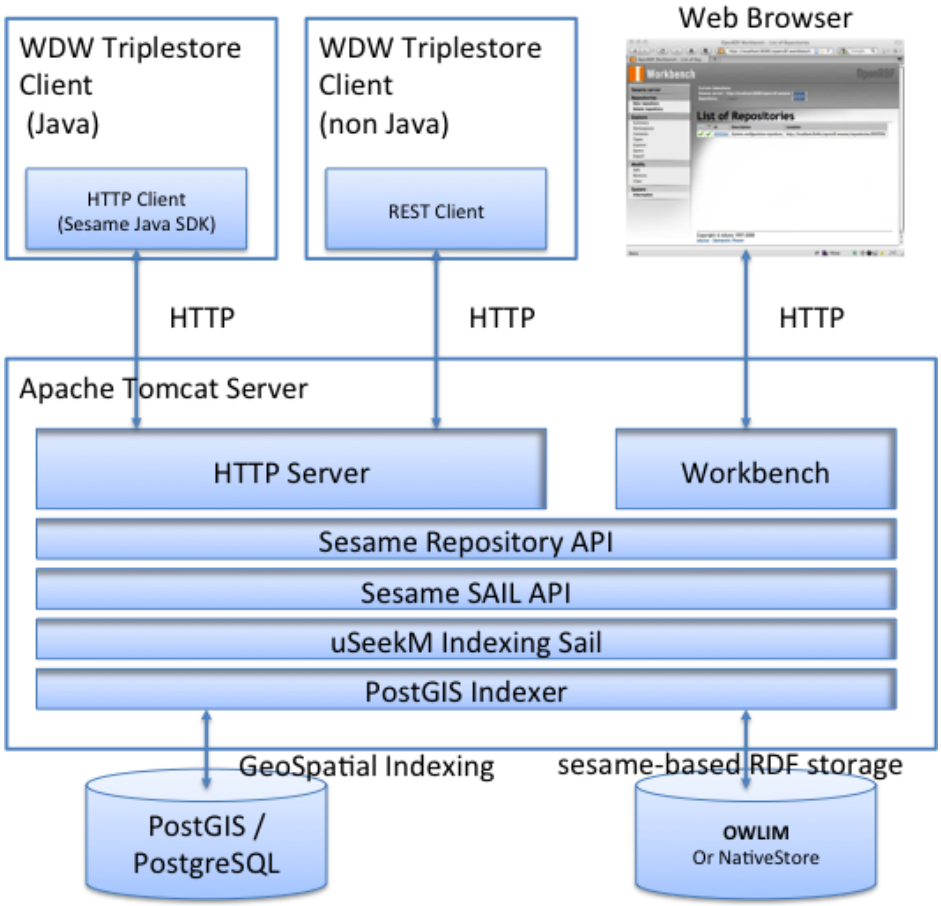

Figure 3: Semantic Reasoning Architecture within the WDW

3.3. ETL Processes and Aggregations

A typical functionality of a Data Warehouse is the provision of so-called ETL processes *(Extract – Transform – Load)* in order to pre-process incoming data, for instance for data cleansing or for data integration. Moreover, it is often useful to create certain (spatial, temporal or topical) data projections, views or aggregations and store them explicitly for more efficient further processing. In WatERP, we employ the R software workbench for realizing such pre-processing algorithms because R offers already a huge number of efficient implementations of statistical and geostatistical data-processing routines. There is also already a connection between R and SOS. Figure 4 indicates how we integrated R routines into the WDW architecture: As soon as an SOS sensor receives new data, this may trigger predefined processing steps. Such a processing step is defined by the R script to be run and its appropriate parameters. The result of the script-application is a modified

[9] http://www.opengeospatial.org/standards/geosparql

[10] http://www.ontotext.com/owlim

measurement-data time-series which is then stored in the SOS server as a "derived SOS sensor", i.e. a virtual sensor.

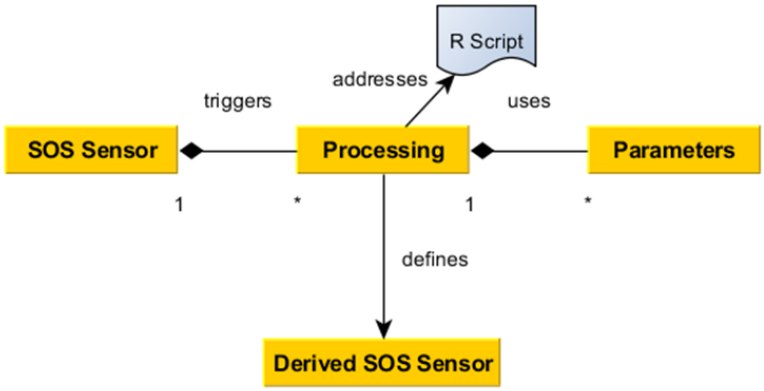

Figure 4: Integration of R Pre-Processing Scripts into the WDW Architecture

4. Conclusions

Status and real-world impact. At the time of preparing this paper, the first fully working prototype of the WDW is up and running and is actually being filled with real-world test data of the two WatERP pilot users, the *Catalan Water Agency* (Agencia Catalana de l'Aigua / ACA, Barcelona, Spain) and the water utility of the city of Karlsruhe (*Stadtwerke Karlsruhe* GmbH, Germany). In the ACA case, e.g., Integrated Water Resource Management is supported such that, through the WDW, the water supply measurements are connected with a Water Demand Management System and a Decision Support System. The interoperation of these different tools and the data exchange between different parties shall help to more efficiently distribute the water resources stored in reservoirs among different utility companies working with one bulk supplier and also among the different competing kinds of water usage (agriculture, industrial use, domestic use).

Although first estimations in the project suggest that water savings in the dimension of 8% in the ACA pilot and energy savings in the dimension of 5% in the Stadtwerke Karlsruhe pilot may be achievable through more integrated and smarter software, it would be dubious to make such promises at this point in time. First, we deliver and test a technical solution. The concrete impact of applying such solutions highly depends on the exact usage situation and the exact way of usage. But we hope to gain more insights about that in the piloting phase 2014/15. In general, the success of many specific use cases will also highly depend on non-technological issues, like the regulatory framework, the incentive systems for data exchange, etc. It might also be the case that for an operational roll-out of a WatERP-like solution, a very strong privacy and data-protection layer would have to be integrated because stakeholders might not send their data to a central system if it was not ensured that unauthorized access is impossible.

Scientific-technical contribution. The main contribution consists in the engineering and the proof-of-concept implementation of the WDW architecture with its different kinds of data and knowledge and different kinds of reasoning possible. This concerns, on one hand, the stability and usefulness of the selected and integrated elements (52°North SOS server, SOS protocol, WaterML2.0 data schema, R server, OWLIM, uSeekM, …) and, on the other hand, the way of integration (separate storage areas for measurement data and semantic knowledge, defining R processing results as derived sensors, …). In general, we believe that such a combined solution of a standards relational data warehouse with a powerful spatial-reasoning enabled technology is unique.

Limitations and next steps. Technically, this prototype is not yet suitable for processing many high-frequency data streams in (near)real-time (for instance, continuous gauge values from thousands of metering points in a flood-endangered river catchment area, or continuous smart-meter data about water consumption of thousands of households). Instead, both pilot application scenarios can live with considering several hundreds of domain objects (water-transmission points between stages in the supply chain, deep-wells, water works, reservoirs or elevated reservoirs, head pipes, pressure zones, …) and a data actualization interval in the dimension between an hour and a couple of days. Scaling-up our solution towards a more real-time and more fine-grained data-collection mode would certainly create new challenges regarding performance and software architecture. Here, recent methods from *Big Data* processing could be employed. Nevertheless, already the current approach offers manifold potential benefits to our end users. It is not obvious whether and when they would technically and organisationally be able to create and reasonable employ huge, high-frequency data streams.

To sum up, the main emphasis of the last project phase in Autumn 2014 – Autumn 2015 will be the piloting in the two test beds. Technically, this could raise some additional work with respect to software usability, stability and efficiency. Also, the integration of the WDW with the complex WatERP overall infrastructure and its deployment in the real-world scenarios will not be trivial. Nevertheless, we are confident that we can provide functional and stable technical prototypes. The next interesting question will be whether the expected benefits for the pilot users can be realized. Within the area of water supply, there are certainly further additional potential benefits from increased data exchange between different parties that were not yet considered in depth. For instance, leakage detection and long-term pipe-network planning could be improved. Further, we expect that the technical solution of WatERP for multi-stakeholder data exchange can be applied far beyond the scope of water resource management, e.g. in the general broader scope of *Smart Cities.*

Acknowledgment*. The work presented here is partially funded by the European Commission under grant EU-FP7-ICT-318603 (WatERP - http://www.waterp-fp7.eu/). We gratefully acknowledge the contributions of Barcelona Digital and Staffordshire University (group of Prof. Wenyan Wu) who also participate in the WDW workpackage of the WatERP project, as well as INCLAM S.A. which greatly supports all software-architectural discussions in the project. Further, we thank the anonymous reviewers for some good ideas and comments.*

References

[1] Anzaldi, G., Chomat, C., Rubión, E., Corchero, A., Moya, A., Moya, C., Ciancio, J. and Helmbrecht, J., "A Holistic ICT Solution To Improve Matching Between Supply And Demand Over The Water Supply Distribution Chain," In: *SDEWES2013, 8th Conference on Sustainable Development of Energy, Water and Environment Systems, Dubrovnik, Croatia,* 2013.

[2] Anzaldi, G., Wu, W., Abecker, A., Rubión, E., Corchero, A., Hussain, A. and Quenzer, M., "Integration of Water Supply Distribution Systems By Using Interoperable Standards To Make Effective Decisions," In: *HIC-2014, 11th International Conference on Hydroinformatics. New York, USA,* 2014. *Forthcoming.*