Computer Science • 13 (1) 2012 http://dx.doi.org/10.7494/csci.2012.13.1.5

Ondrej Habala
Ladislav Hluchý
Viet Tran
Peter Krammer
Martin Šeleng

USING ADVANCED DATA MINING AND INTEGRATION IN ENVIRONMENTAL PREDICTION SCENARIOS

Abstract *We present one of the meteorological and hydrological experiments performed in the FP7 project ADMIRE. It serves as an experimental platform for hydrologists, and we have used it also as a testing platform for a suite of advanced data integration and data mining (DMI) tools, developed within ADMIRE. The idea of ADMIRE is to develop an advanced DMI platform accessible even to users who are not familiar with data mining techniques. To this end, we have designed a novel DMI architecture, supported by a set of software tools, managed by DMI process descriptions written in a specialized high-level DMI language called DISPEL, and controlled via several different user interfaces, each performing a different set of tasks and targeting different user group.*

Keywords data mining, data integration, meteorology, hydrology

1. Introduction

Modern society with ever-evolving methods of fast transportation, expanding urban centers and increasing population density in previously unpopulated rural areas requires always accurate meteorological predictions not only of general weather conditions, but also of various significant meteorological phenomena [1]. For some of these, there are no accurate physical models, or if they are available, their customization to a particular target area is unfeasible because of its complexity and often missing prerequisites, like past observations or a detailed topological map. To overcome these difficulties, we have performed several experiments by applying data mining techniques to a set of carefully chosen meteorological and hydrological scenarios. While data mining has been used in meteorology for a long time, the scenarios we have chosen have not been previously covered, especially not in the target area we have chosen. They have been designed and evaluated by domain experts, and their design was driven by the current needs of these experts and their employers. In their design we have also used our previous experience in applying information technologies to environmental predictions [6]. These experiments are part of the FP7 project ADMIRE[1], and additionally to serve as an experimental platform for meteorologists and hydrologists, we have used them as a testing platform for a suite of advanced data integration and data mining (DMI) tools, developed within this project. The idea of the project ADMIRE is to develop an advanced DMI platform accessible even to users who are not familiar with data mining techniques. To this end, we have designed a novel DMI architecture [5], supported by a set of software tools, managed by DMI process descriptions written in a specialized high-level DMI language called DISPEL [2], and controlled via several different user interfaces, each performing a different set of tasks and targeting different user group. In this paper we present the results of the project ADMIRE from the point of view of our environmental pilot application. We describe the methods ADMIRE uses to integrate geographically distributed data sets, stream them through a series of filters and processing elements using the OGSA-DAI platform [9], and deliver the results to the end users who have requested them. The project has successfully finished with a final review in July 2011, and the final platform allows for the easy development of complex DMI scenarios using an existing library of processing elements.

2. Data Intensive Processes in the Environmental Domain

Environmental risk management research is an established part of the Earth sciences domain, already known for using powerful computational resources to model physical phenomena in the atmosphere, oceans and rivers [15]. In this paper we show how state-of-the-art tools for data-intensive processes can be applied to the benefit of

[1] ADMIRE – Architectures for Data Intensive Research. http://www.admire-project.eu/

meteorological and hydrological experts. We illustrate the possibilities on a simple scenario from the hydro-meteorological domain.

The environmental domain makes extensive demands of data management as well as data processing. A significant portion of data input to weather prediction models (both physical and statistical) comes from observations. Only a fraction of these observations can be made remotely (satellite and radar observations are examples of remote weather sensing), and the bulk of the input data comes from local observations of air temperature, humidity, pressure, precipitation, and other parameters. Local observations tend to produce local data sources, and a large number of local observations leads to a large number of local data sources, and a significant level of distribution of the data. Even in a context where there is a national weather management authority, several geographically dispersed data sources are needed for any integrated analysis. For example, in Slovakia we might consider these data sources:

- meteorological observation database owned by the Slovak Hydrometeorological Institute (SHMI)[2];
- weather radar observations conducted by the Slovak Hydrometeorological Institute, and stored in a separate data store;
- hydrological observations conducted by branches of Slovak Water Enterprise (SWE)[3], and stored locally at the respective branches;
- waterworks manipulation schedules, local to the management centre of the respective waterworks (there are four such centres);
- local observations by specialized personnel at airports and other installations, where weather is a major operational concern.

This list is not complete and the number of sources may be much larger. A traditional approach to data management in this environment is to establish a list of necessary data sources, perform negotiations with their owners, acquire the data (usually in a form of a static database image), assess its quality, prepare it, and then feed it to the model. This process is cumbersome, can take months and is not suited for day-to-day weather management operations.

A different approach is to use modern methods of distributed data management: establish a Quality of Service agreement, an on-line data integration process including all necessary data preparation and filtering, and make the process as automated as possible. The main difference is in treating the input data not as a suite of static data sets but rather *as a group of converging data streams*. Weather does not cease to exist at the moment a snapshot ends. While this approach is still a novelty for domain experts trained in a different context and accustomed to data transfer channels with much smaller bandwidths, it has been recognized as the future of environmental data distribution, as can be seen in the pan-European INSPIRE Directive [4], man-

[2]Slovenský hydrometeorologický ústav, Jeséniova 17, 833 15 Bratislava, Slovakia, `http://www.shmu.sk`

[3]Slovenský vodohospodársky podnik, štátny podnik, Radničné námestie 8, 969 55 Banská Štiavnica, Slovakia, `http://www.svp.sk`

Table 1

Nature, current size and the approximate rates of increase of the various data sources used in the environmental risk management scenarios. While training the various predictive models makes use of historical snapshots, the live system is designed to work with incoming streams of data from the various sources, some of which change slowly but significantly over the course of a year

Data set	Source	Nature	Size (MB)	MB/year
NWP data	SHMI	Simulation	arbitrary†	c. 20,000
Synoptic stations	SHMI	Sensor	50	1
Rainfall measurement	SHMI	Manual	100	< 1
Radar imagery	SHMI	Sensor	10,000	300
Waterworks	SWE	Manual	300	20
Hydrology stations	SHMI	Sensor	300	30

† can be regenerated as required

dating the use of service-oriented architectures and a whole suite of standards for environmental data publication and access.

The experiments we've performed are dependent on several input datasets, which we summarize briefly in Table 1. We measure these datasets not only in terms of absolute size but also in terms of how quickly they are increasing. While many of the datasets comprise a large part a historical corpus of measurements, new observations and new simulations are being performed and added all the time. Training of any given predictive model [8] will, of course, make use of historical data, but the actual *use* of such a model will require the most up-to-date data available. This is a good illustration of the concept of "thinking in streams" – data are dynamic, not just static blocks.

3. The ORAVA Scenario — Mining for Water Level and Temperature

This scenario was defined by the Hydrological Service division of SHMI. Its goal is to predict the water discharge wave and temperature propagation below the Orava reservoir, one of the largest reservoirs in Slovakia [3, 7].

This scenario covers a relatively small area of northern Slovakia (see Figure 2). The selected data which influence the scenario's target variables – the discharge wave propagation and temperature propagation in the outflow from the Orava reservoir to the Orava river – are shown in Table 2. The data are gathered from the hydro-meteorological sensor networks of several data providers. Figure 1 shows the layout of the sensors below the Orava reservoir. Orange dots represent the sensor network of SWE, which provide reservoir water temperature and discharge data. Red dots show a part of the network of hydrological sensors operated by SHMI. These sensors are stationed in the Orava river and its

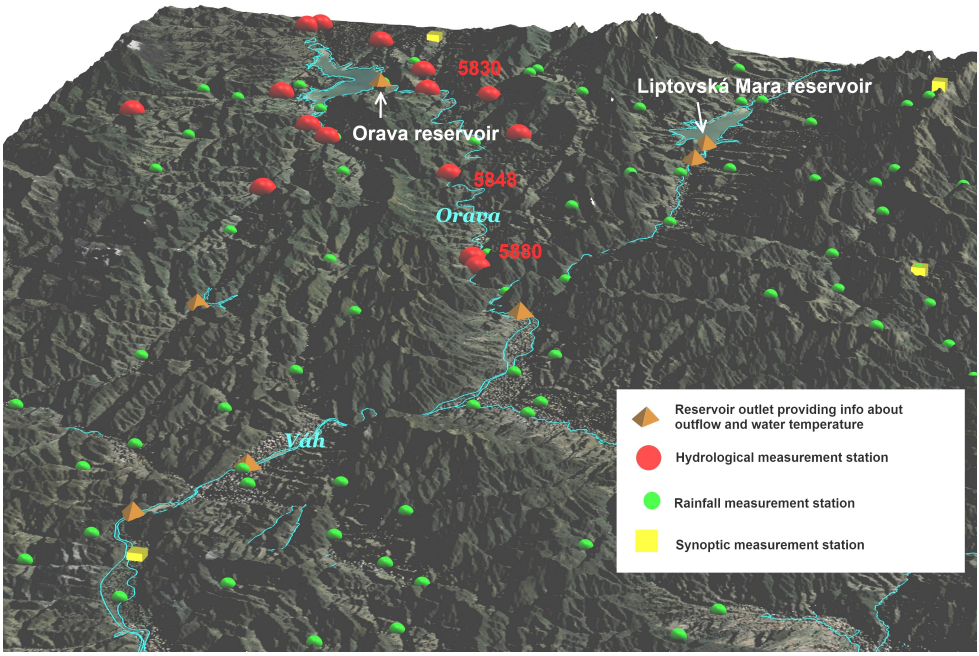

Figure 1. A visualization of an actual network of hydro-meteorological sensors in the northern part of Slovakia, around the Orava reservoir

tributaries, and measure current river temperature and water level. The densest sensor network, depicted by green dots, is the network of precipitation measurement stations, providing hourly precipitation data. Additional to these, there are more complex synoptic sensor stations, depicted in yellow, which provide precipitation and other climatological measurements. What is not shown in the picture is the mesh of values provided by meteorological radar and meteorological simulations.

As predictor variables in this scenario (shown in Table 2), we have selected rainfall and air temperature, the discharge volume of the Orava reservoir and the temperature of water in the Orava reservoir. Our target variables are water height and water temperature measured at a hydrological station below the reservoir. As can be seen in Figure 2, the station directly below the reservoir is number 5830, followed by numbers 5848 and 5880 – these stations are the target sites for which predictions are made. If we run the data mining process at time t, we expect to know all sensor data up to this time (first three data lines in Table 2). The future rainfall and temperature values are obtained by running a standard meteorological model. Future discharge rate of the reservoir is given in the management schedule of the reservoir. The actual data mining targets are the X and Y variables for times after time t.

Table 2

Schematic depiction of the predictor variables and targets in the water level and temperature prediction scenario. T_A denotes air temperature, T_R reservoir temperature and T_{St} the temperature at the water station in question. H_{St} is the station height above sea level

Time	Rainfall	T_A	Discharge	T_R	H_{St}	T_{St}
$t-2$	R_{T-2}	F_{T-2}	D_{T-2}	E_{T-2}	X_{T-2}	Y_{T-2}
$t-1$	R_{T-1}	F_{T-1}	D_{T-1}	E_{T-1}	X_{T-1}	Y_{T-1}
t	R_T	F_T	D_T	E_T	X_T	Y_T
$t+1$	R_{T+1}	F_{T+1}	D_{T+1}	E_{T+1}	X_{T+1}	Y_{T+1}
$t+2$	R_{T+2}	F_{T+2}	D_{T+2}	E_{T+2}	X_{T+2}	Y_{T+2}

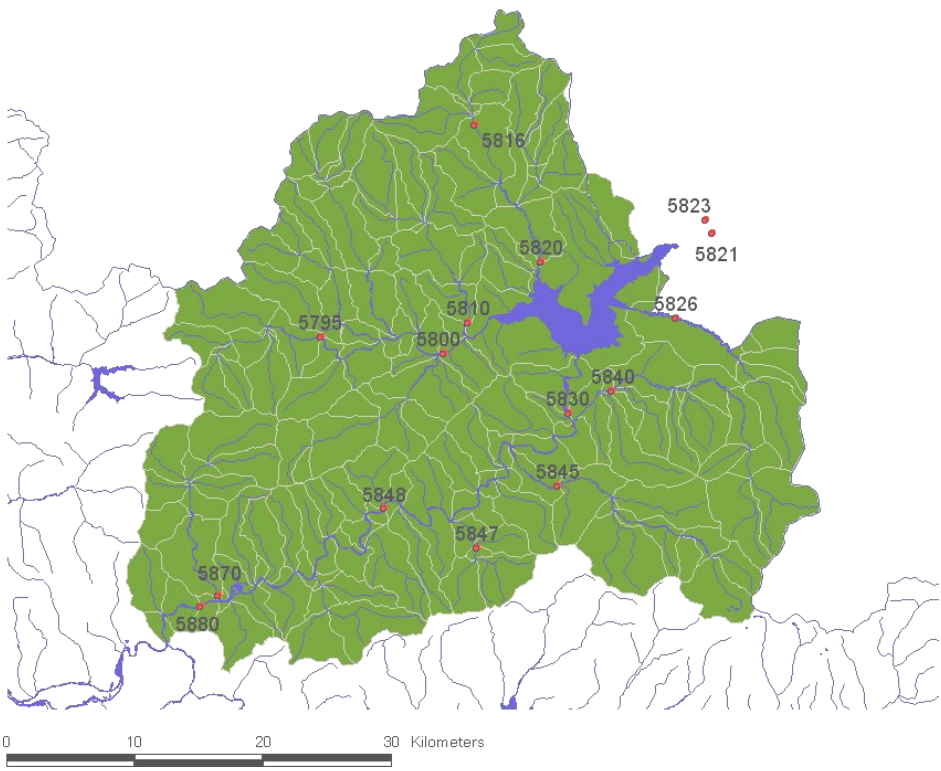

Figure 2. The geographical area of the ORAVA pilot scenario

3.1. Management of Distributed Data Mining and Integration

The described scenario can be divided into three processes: data integration, training, and prediction.

The first process, shown in Figure 3, integrates required data from the distributed data sources and saves the result to a file repository in the form of a stream of tuples.

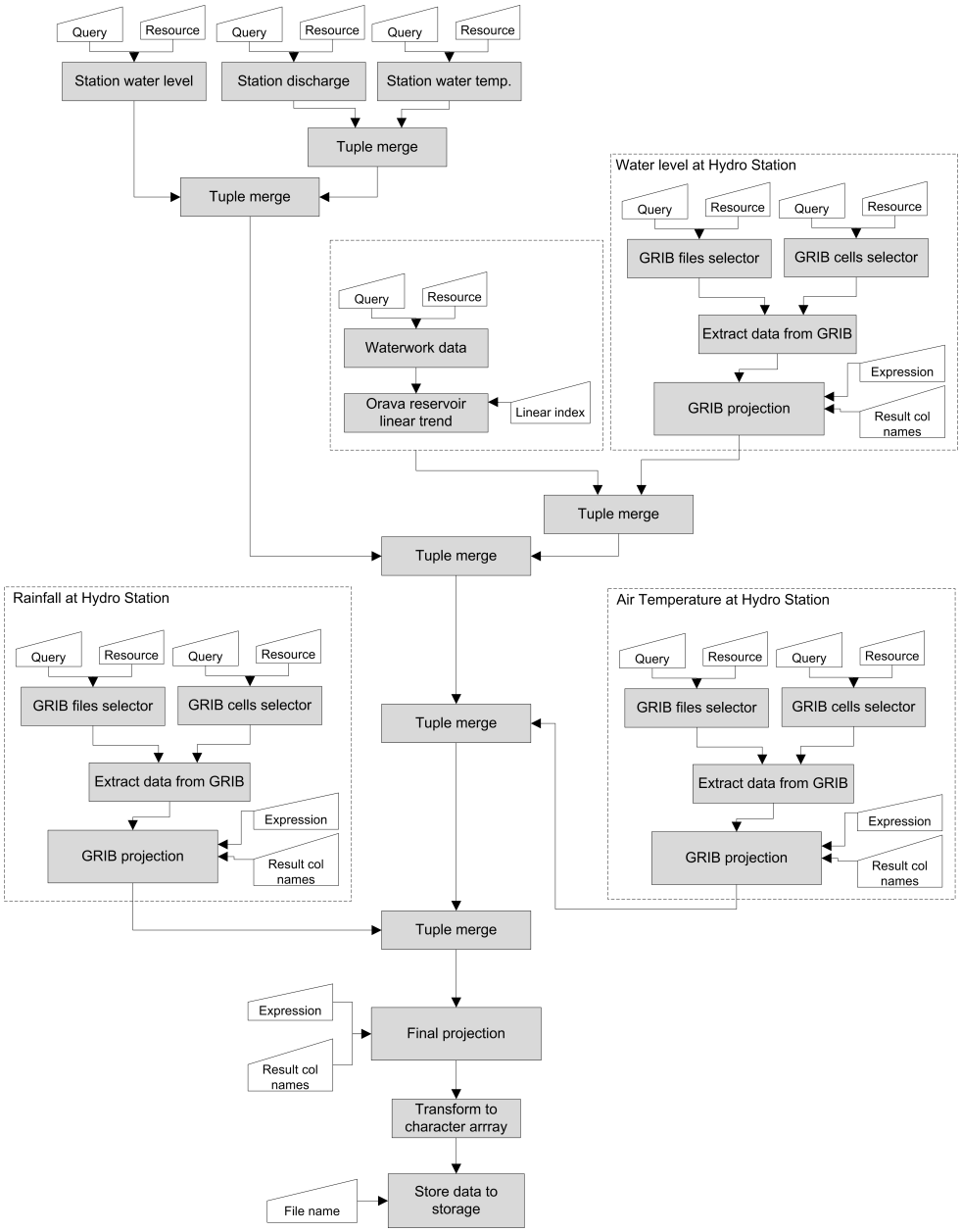

Figure 3. A graphical representation of the data integration process in the ORAVA hydrological scenario predicting water level and water temperature

The process begins by extracting data about the relevant hydrological measurement station from a relational database (station water level, station discharge, station water temperature). These values are merged into the initial tuple, which is then expanded further as the process progresses by means of a simple `Tuple merge` operation (employed numerous times throughout the process). In parallel we read data from the reservoir database (operated by a separate entity), and also access various parameters present in the computed weather data stored in the form of GRIB (GRIdded Binary) [13] files.

The data from the reservoir need to be filtered by the operation `Orava reservoir linear trend`, which replaces missing values in the data by a linear interpolation. Also, any reading of a particular GRIB file is preceded by access to the GRIB metadata database, which holds information about the contents of the GRIB files. After integrating all of these data into a wide tuple, the tuple is filtered to remove duplicate occurrences of some parameters (for example the date and time, which are used by `Tuple merge` to synchronize the data stream), and the result is stored in a file repository for later use.

The second process, not shown here in graphical form, reads the integrated data from the stored file, deserializes it and builds a linear regression classifier using parameters set and verified by a data mining expert. The trained model is then serialized and stored back in the repository.

The final part of this scenario is the actual prediction process. This process expects to find already-integrated data in the file repository, as created by the integration process (Figure 3). It also downloads the trained data mining model, feeds the integrated data into the model, and then merges the original data with the predicted ones. This process may be used for both verification of the model (if we use past data for prediction) and for the actual prediction (if we use data containing future weather prediction).

This scenario illustrates very well the complexity of the data preparation and integration stages of the analysis process. The design of the model algorithm is actually a rather small part of the overall knowledge discovery process.

4. New technologies for environmental data mining

The DISPEL language allows us to describe the processes in each of the three environmental risk scenarios at a high level of abstraction, independently of any low-level concerns regarding the underlying enactment engines, databases or any consideration of the distributed environment. Our experiments have made use of several interconnected gateways, which together provide all the necessary data, processing elements and visualization tools which our scenarios require. A sample of the source code for the data integration portion of our scenario is in Figure 4. Here we can see the main building blocks of a DISPEL program. After initial declarations of the processing elements `grib`, `gribFileSelector` and `gribCoordsSelector`, and the string literals `gribID`, `gribFileSelectQuery` and `gribCoordSelectQuery` (the latter two being obviously SQL

queries used to access a GRIB metadata repository), we can see the construction of a simple stream in a line reading:

```
|- gribFileSelectQuery -| => gribFileSelector.expression;
```

The operators `|-` `-|` denote the beginning and end of a stream of literals – in this case just one literal, the SQL query used to select the proper GRIB file. The operator `=>` then creates a connection of this stream in an input connector (called `expression`) of the processing element `gribFileSelector`. This way we feed a stream into a processing element (PE), and we can also connect an output connector of one PE to the input of another PE. So these two operators allow us to create a graph in which the data is streamed and processed by various PEs. For a more thorough explanation of the language please consult the User Manual [11][4].

```
package eu.admire.demo.orava.integration {


use uk.org.ogsadai.SQLQuery;
use uk.org.ogsadai.TupleToWebRowSetCharArrays;
use eu.admire.spatioTemporal.activities.tupleMerging.OrderedTuplesMerge;
use eu.admire.spatioTemporal.activities.grib.GribCellValues;
use uk.org.ogsadai.TupleSimpleMerge;
use eu.admire.Results;
use eu.admire.LinearTrendFilter;
use uk.org.ogsadai.TupleArithmeticProject;
use eu.admire.BuildClassifierLinearRegression;
use eu.admire.Classify;
use eu.admire.Serialiser;
use uk.org.ogsadai.DeliverToFTP;
use uk.org.ogsadai.TupleToCSV;


// Composite PE for reading data from Grib.
// Parameters:
// startDate, endDate: start and end date of duration, in form yyyy-mm-dd
// value: code of data from grib (CWDI for precipitation and TMP for temperature)
// longitude, latitude: longitude and latitude of selected location (Orava waterwork or water stations along river)
PE(<> => <Connection output>) readingGribFunction(String startDate, String endDate, String value, String longitude,
    {
        GribCellValues grib = new GribCellValues;
        SQLQuery gribFileSelector = new SQLQuery;
        SQLQuery gribCoordsSelector = new SQLQuery;

        String gribID = "DbGribMetaResource";

        String gribFileSelectQuery = "SELECT file, date_time, type, type_desc, type_unit FROM grib_meta_c WHERE type
            and   date_time >='" + startDate + " 00:00:00' and date_time <='" + endDate + " 24:00:00' and forecast=0
        String gridCoordSelectQuery = "SELECT id AS ID, lat, lon, ( sqrt( ((lon-" + longitude + ")*(lon-" + longitud
            ((lat-" + latitude + ")*(lat-" + latitude + ")) ) ) as dist FROM `grid_coords2` order by dist limit 0,1;

        |- gribFileSelectQuery -| => gribFileSelector.expression;
        |- gribID -| => gribFileSelector.resource;

        |- gridCoordSelectQuery -| => gribCoordsSelector.expression;
        |- gribID -| => gribCoordsSelector.resource;

        gribFileSelector.data => grib.gribFiles;
        gribCoordsSelector.data => grib.gribCells;

        return PE (<> => <Connection output = grib.gribResult> );
    }
```

Figure 4. A sample of DISPEL code of the ORAVA scenario

This novel approach allows us easily to extend the hydro-meteorological infrastructure to new data providers, by deploying a gateway at the site of the new provider,

[4] `http://www.admire-project.eu/docs/DISPEL-manual.pdf`

and registering it with the other gateways. Then, when a data analysis expert creates a DISPEL document that makes use of one of the capabilities provided by this gateway, it can be accessed and integrated automatically into the overall knowledge discovery workflow.

This model provides a clear separation of responsibilities between data-intensive engineers, data analysis experts, and the domain experts of the application. The underlying infrastructure and gateway network is managed by the data-intensive engineers. The data-analysis experts use DISPEL to create full knowledge discovery workflows which *utilize* the infrastructure without needing to *understand* it. In turn these workflows are used by the domain experts via specialized domain-specific portals.

This approach also provides for a reasonable amount of fault tolerance. If one data centre becomes unavailable, it may conceivably be replaced transparently by a different one, without the final users of our product ever knowing it happened. Some centres and gateways are, of course, irreplaceable in a given network (primary data storage centres for instance), but data filters may be deployed at several locations to enhance redundancy. There may also be several HPC facilities available to a given user, so the temporary inaccessibility of one of them is no issue – the DISPEL description of the required data-oriented solution is entirely agnostic of such things.

5. Conclusion

We have presented an approach to hydro-meteorological predictions, which utilizes state-of-the-art data integration and data mining technology developed in the EU FP7 project ADMIRE. The technology allows us to make the whole DMI process much more flexible, and especially the DISPEL language developed in the project makes the experience much more pleasant even for novices to data mining. Once the DMI process is described in a DISPEL program, it can be reused even when the underlying middleware and hardware configuration changes significantly. It is based on the popular framework OGSA-DAI, which enjoys good support by its authors and an extensive user base.

We continue to use the results of ADMIRE even after the project has finished. The concept has proven its usefulness and we are already working on porting other application scenarios to the ADMIRE framework, as can be seen for example in [1]. Also there are other works which already utilize ADMIRE technology for different application domains [14, 12, 10].

Acknowledgements

This work is supported by projects ADMIRE FP7-215024, APVV DO7RP-0006-08, DMM VMSP-P-0048-09, Project ITMS: 26240220029, ITMS: 26240120029, VEGA No. 2/0054/12.

References

[1] Bartok J., Habala O., Bednar P., Gazak M., Hluchy L.: *Data Mining and Integration for Predicting Significant Meteorological Phenomena*. [in:] Proc. of ICCS 2010 — International Conference on Computational Science, vol. 1 of *Procedia Computer Science*, pp. 37–46. Elsevier Science BV, 2010.

[2] Brezany P., Aranda C. B., Corcho O., Janciak I., Woehrer A., Atkinson M.: *ADMIRE – Report Defining the Final Iteration of the Model and Language*. Deliverable report D1.9, The ADMIRE Project, May 2011.

[3] Ciglan M., Habala O., Tran V., Hluchy L., Kremler M., Gera M.: *Application of ADMIRE Data Mining and Integration Technologies in Environmental Scenarios*. [in:] R. Wyrzykowski, J. Dongarra, K. Karczewski and J. Wasniewski (Eds.), Parallel Processing and Applied Mathematics, Part II, volume 6068 of Lecture Notes in Computer Science, pp. 165–173. Springer-Verlag Berlin, 2010.

[4] EU Parliament: *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*. Official Journal of the European Union, 50(L108), April 2007.

[5] Galea M., Atkinson M., Liew C. S., Martin P.: emphADMIRE – Final Report on the ADMIRE Architecture, May 2011.

[6] Habala O., Mališka M., Hluchý L.: *Service-based flood forecasting simulation cascade in k-wf grid.* [in:] M. Bubak, S. Unger (Eds.), Cracow 06 Grid Workshop : K-Wf Grid, pp. 138–145. Academic Computer Centre CYFRONET AGH, 2007.

[7] Hluchy L., Habala O., Tran V., Ciglan M.: *Hydro-meteorological scenarios using advanced data mining and integration*. [in:] *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*, volume 7, pp. 260–264, aug. 2009.

[8] Hluchý L., Šeleng M., Habala O., Krammer P.: *Mining Environmental Data in Hydrological Scenarios*. [in:] M. Li, Q. Liang, L. Wang, Y. Song (Eds.), *Seventh International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010, 10-12 August 2010, Yantai, Shandong, China*, pp. 1988–2992. IEEE, 2010.

[9] Jackson M. J., Antonioletti M., Dobrzelecki B., Hong N. C.: *Distributed data management with ogsadai*. [in:] S. Fiore, G. Aloisio (Eds.), Grid and Cloud Database Management, pp. 63–86. Springer Berlin Heidelberg, 2011.

[10] Jarka M., Podraza R.: *Architecture of distributed system for challenging data mining tasks*. [in:] D. Ryzko, H. Rybinski, P. Gawrysiak, M. Kryszkiewicz (Eds.), Emerging Intelligent Technologies in Industry, volume 369 of *Studies in Computational Intelligence*, pp. 197–206. Springer Berlin / Heidelberg, 2011.

[11] Martin P., Yaikhom G., et al.: *Dispel: Data-intensive systems process engineering language, user manual*. Technical report, August 2011.

[12] Spinuso A., Trani L., Atkinson M., Galea M.: *Infrastructure for data-intensive seismology: Cross-correlation of distributed seismic traces through the admire*

framework Geophysical Research Abstracts, 13, 2011.

[13] Stackpole J.: *The WMO format for the storage of weather product information and the exchange of weather product messages in gridded binary form.* U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, National Meteorological Center, 1994.

[14] Trani L., Spinuso A., Galea M., Main I.: *A novel automated approach to ambient noise data processing using the admire framework. Geophysical Research Abstracts*, 13, 2011.

[15] Yokokawa M., Itakura K., Uno A., Ishihara T., Kaneda Y.: *16.4-tflops direct numerical simulation of turbulence by a fourier spectral method on the earth simulator.* SC Conference, 0:50, 2002.

Affiliations

Ondrej Habala
Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia,
ondrej.habala@savba.sk

Ladislav Hluchý
Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia

Viet Tran
Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia

Peter Krammer
Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia

Martin Šeleng
Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia

Received: 6.12.2011
Revised: 17.01.2012
Accepted: 30.01.2012