International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Fusion of Environmental Information for the Delivery of Orchestrated Services for the Atmospheric Environment in the PESCaDO project

Victor Epitropou[1], Lasse Johansson[2], Kostas D. Karatzas[1], Anastasios Bassoukos[1], Ari Karppinen[2], Jaakko Kukkonen[2], Mervi Haakana[2]

[1] *Aristotle University of Thessaloniki, Dept. of Mechanical Engineering, Informatics Systems and Applications Group, 54124 Thessaloniki, Greece.*

[2] *Finnish Meteorological Institute, P.O. Box 503, FI-00101, Helsinki, Finland*

(vepitrop@isag.meng.auth.gr, Lasse.Johansson@fmi.fi, kkara@eng.auth.gr, abas@isag.meng.auth.gr, Ari.Karppinen@fmi.fi, jaakko.kukkonen@fmi.fi, Mervi.Haakana@fmi.fi)

Abstract: The PESCaDO project (http://www.pescado-project.eu/) aims at providing tailored environmental information to EU citizens. For this purpose, PESCaDO delivers personalized environmental information, based on coordinating the data flow from multiple sources. After the necessary discovery, indexing and parsing of those sources, the harmonization and retrieval of data is achieved through Node Orchestration and the creation of unified and accurate responses to user queries by using the Fusion service, which assimilates input data into a coherent data block according to their imprecision and relevance in respect to the user defined query. Environmental nodes are selected from open-access web resources of various types, and from the direct usage of data from monitoring stations. Forecasts of models are made available through the synergy with the AirMerge Image parsing engine and its chemical weather database. In the presented paper, elements of the general architecture of AirMerge, and the Fusion service of PESCaDO are exposed as an example of the modus operandi of environmental information fusion for the atmospheric environment.

***Keywords**:* environmental information fusion, environmental image reverse engineering service.

1 INTRODUCTION

In recent years, the emergence of social media, personalized web services and an increased public awareness of environmental factors that impact Quality of Life (QoL) have led to the demand for easier access to environmental information and its tailoring to personal needs. According to Klein et al. [2012] and Karatzas and Kukkonen [2009], there is a need for an integrated assessment of the impact of air pollution, allergens and extreme meteorological conditions on public health.

Chemical weather is defined as the short-term (less than two weeks) variability of the atmospheric chemical composition, as per Kukkonen et al. [2012]. Recently, a new open-access CW forecasting service has been set up, which is user-friendly, dedicated to the whole European continent, and which can automatically extract data from various CW sources as described in Balk et al, [2011]. While there are

"web mashup" services such as the ones described by Ganesh et al. [2004] that combine the ability to use data feeds and human-readable articles as data sources for building new web services, they are limited to textual data harvesting and replication (for example, by linking and reproducing existing articles). In the CW domain, it is currently not trivial, for example, to get a direct answer to seemingly simple questions like "How will the air quality be tomorrow in Helsinki?" without a lot of manual search and interpretation of the often contrasting information found on various web sites, even if proper and up-to-date data sources are available.

The purpose of the PESCaDO project is to address the above point, by (i) taking as a starting point plain text queries in human languages or interactive user input through a web interface, in a way that transcends simple keyword-based searches (ii) infer the implied context and semantic meaning of the user's query (e.g., place, time, activity, type of information desired) by using advanced semantic and ontological textual analysis tools, and (iii) coordinate the data flow from a number of heterogeneous (text, images, feeds, binary files) data sources (Environmental Node Orchestration) in order to produce a response by taking into account all available data, as described in Wanner et al. [2011].

2 ELEMENTS OF PESCADO'S ARCHITECTURE

The PESCaDO service system currently includes modules to perform the following tasks: i) discovering new data sources and assess their reliability and historical performance, ii) integrating any newly acquired data into the system as new processed ("cooked") information, iii) being able to retrieve this information afterwards, iv) performing fusion of the archived data and v) present any results to the end user in an understandable form. In particular, the PESCaDO system also extracts information from the so-called "Invisible Web", as defined by Bergman [2001]. With invisible web, we refer to the web-based data that are hard to find due to their unsupported data formats and their deep layering. The invisible web sources are normally intractable by conventional search engines (e.g. Google).

It is, therefore, a Service Oriented Architecture, consisting of multiple networked service modules, most of them fully web-enabled, each with a specific task and interconnected through mutually agreed upon protocols. The protocols used for internal communication within the various service modules in the PESCaDO project are for the most part specifically designed according to RESTful service guidelines as described by Rodriguez [2008]. A demonstration of the working prototype of PESCaDO is available at http://www.pescado-project.eu/.

Some of the basic service modules used in PESCaDO can be summed up, in terms of functionality, in the following principal types (more details may be found in Wanner et al. [2010], and Wanner et al. [2011]):

- The Node Discovery Service (NDS), tasked with discovering new data sources through link hopping and web crawling techniques
- The Data Retrieval Service (DRS), which retrieves any data stored by the Node Discovery Service (or added to PESCaDO's data pool by other means), and makes it available to computational modules which can use them directly.
- The Fusion Service, which is a specialized module tasked in the assimilation of meteorological and air quality input data by utilizing imprecision metrics, statistical models and optimization methods, and in production of unified and relevant datablocks in respect to the user defined query.
- Auxiliary/Special Purpose nodes or modules, not necessarily developed within and belonging to PESCaDO. One such module is the AirMerge system first described in Epitropou et al. [2010].

A very important requirement in any system based on data from the Web is the ability to filter it, separate the useful content from noise (defined here as irrelevant

content), and be able to interpret it correctly. A second consideration is that after the useful data has been retrieved and parsed, it must be stored and retrieved in a commonly interoperable format, at least within the system itself. Furthermore, the data itself must only be stored, if it actually leads to new knowledge, and must be sufficient in quality and quantity in order for the system to be able to answer user queries with reasonable precision. These tasks are performed to different extents in all of the above subsystems.

2.1 Environmental Nodes

The data sources usable for the purposes of PESCaDO can be distinguished into two broad categories: direct and indirect sources. Direct sources represent web pages (including especially websites that present CW, weather and pollen model results), websites with air quality bulletins, reports or streaming RSS feeds, etc. whose primary purpose is the publishing of environmental information, in particular with regards to the concentration of pollutants, atmospheric conditions, temperature, and so on. Usually this kind of sources is owned or managed by research or other official institutions. The data from the aforementioned direct sources can in turn be replicated or referenced from secondary websites, e.g. information or news portals, which then become secondary sources of information, but are still considered direct sources because the data is presented as-is.

In addition to direct sources of environmental information, there are also implicit (or tertiary) sources which refer to environmental information “between the lines”: such information can be found in media such as amateur websites, discussion forums, blogs and social networks in general. However, this information is not currently included in the PESCaDO system, though methods for their integration are under investigation.

Both in the case of official text-based sources and in the case of indirect sources, ontological-semantic text analysis techniques can be used in order to automatically separate potentially useful content from the plethora of text, as indicated by Ganesh et al. [2004]. In addition to these two distinctions, the following two classes of environmental nodes can be distinguished: textual nodes, and mixed content nodes. The former contain useful data exclusively or mainly in textual form, regardless of its formatting or presentation, and is the typical website scenario. The latter type of nodes contain a significant fraction (or even all) of their environmental information in non-textual forms. These may include, but are not limited to: static images, binary documents, executable files, custom website plugins, videos, audio files etc. With the appropriate modules, PESCaDO is currently able to handle some of those types, in particular static images.

2.2 Environmental Node interconnection

Environmental information in the form of human-readable text or machine-readable, text-format data feeds is a readily available and easily processed resource. However, there are a few types of mixed content nodes, in particular those containing images such as charts, graphics and concentration maps that can be treated as alternative sources of information after specialized processing. The PESCaDO system handles such sources through the use of specific auxiliary modules. This domain-specific capability sets PESCaDO apart from most general-purpose search engines.

Mixed data sources are usually of higher quality, quantity and precision in terms of contents. For example, the information that can be extracted from a single coverage map is equivalent to thousands of data points with clear-cut numerical values and an extended, dense geographical coverage On the contrary, the

amount of information which can be extracted from human-readable text alone will clearly be more limited, even in the case of raw data size parity.

However, identifying and correctly parsing such data sources remains a largely manual process which requires human intervention and overseeing in several phases and usually requires writing ad-hoc auxiliary connector modules that can perform the necessary image and signal processing. Such sources also require human overseeing in order to properly configure them (e.g. compiling a list of reliable and robust sources, and writing custom software which can handle this very specific task). Therefore, PESCaDO uses both approaches, with text-based crawling being built-in the Node Discovery Service and handling only text-based information, and using external services whenever the integration of mixed-type nodes is deemed necessary. Currently, PESCaDO uses a specially designed connector with the AirMerge system, presented in Epitropou, [2011], for the purpose of including the information contained in the CW models of the European Open-access Chemical Weather Portal into PESCaDO's infrastructure.

2.3 Node Orchestration Service

The Node orchestration in the PESCaDO system includes:

- Infrastructure for the interconnection of environmental services,
- Assessment of uncertainty metrics,
- Fusion of data from different sources. and
- Ontology-based service connecting strategy

The concept of node orchestration refers to a coordinated process which includes the interpretation of user requests into a unique spatial-temporal-environmental concern, and the coordination of all available data sources (or "nodes"), using all available services, in order to assemble a response. This service must be able to understand user requests and broaden or restrict search scopes and context as necessary.

As a practical example, a request for the general air quality index in Helsinki, could be schematized as follows: i) the user asks literally "what is the air quality in Helsinki tomorrow" ii) The text analysis identifies both the locus ("Helsinki"), the time ("tomorrow", relative to the day of the request) and the type of request ("air quality", with no other qualifiers). Those clues are rather vague, since "Helsinki" is a geographical region, "tomorrow" can mean a wide range of time instances, and "air quality" can also be defined in numerous ways. These ambiguities are resolved using an adaptive user interface such as the ones described in Judah et al. [2009], which attempt to infer what the user most probably meant, for example, by using information from the user's profile such as habits, location etc. to restrict the subset of possible answers to a context that is more likely to be close to the user's needs. This strategy is increasingly used in intelligent software agents such as Siri from Apple, Inc.

Given the above, all available data sources that cover the area of Helsinki and other nearby municipalities are queried for data and refer to any time instance on the day described by "tomorrow" relative to the request's timestamp. These can range from one-word remarks e.g. "good" or be detailed concentration values for specific pollutants on relatively narrow geographical boundaries (e.g. on 1km x 1km boundaries). After fusion, the user can then be presented with a summary of all relevant information, which can be either direct e.g. in the form of official air quality indexes based on EU regulations, or computed from the synergy of several sources (e.g. using formulas to compute the impact of different pollutants to air quality).The coordinated use of all available data sources and environmental knowledge stored into the PESCaDO system is the essence of orchestration.

2.4 Fusion Service

The Fusion Service is a special PESCaDO data module, in that it acts as a processor for the data retrieved by the DRS, and as such it has a dual data source-processing module nature. Despite this, it is considered an integral component of PESCaDO.

In the fusion process, all pieces of meteorological data and air quality data are considered to reflect conditions at a certain time (t) and place ($\mathrm{r} = [latitude, longitude]$). These pieces of information are regarded as estimators $\hat{\theta}_i(\mathrm{r_i}, t_i)$ for the conditions $\varphi(\mathrm{r_0}, t_0)$ in the user defined area and time:

$$\varphi(\mathrm{r_0}, t_0) = \hat{\theta}_i(\mathrm{r_i}, t_i) + \varepsilon$$

where ε is the error in terms of statistical variance and bias (e.g. unreliable source or irrelevant data). The fusion service estimates an aggregate statistical variance measure for each piece of input data and the derived imprecision metrics are then used for input data weight assignment. Essentially, a large estimated aggregate imprecision measure causes the assigned weight to decrease while the data from the more accurate and relevant sources gain more emphasis in the fusion [Potempski et al, 2008].

The total imprecision models, which have been formulated individually for each air pollutant type using regression analysis with historical data, estimate a statistical variance measure for (i) temporal separation component, (ii) spatial separation component and (iii) input source base variance component. For the latter, the fusion service utilizes a historical input source performance database which stores information about the known data node's prediction accuracy in the past. A schematic diagram of this total imprecision estimation process is presented in Figure 1.

The three individual components described in the total variance model can be regarded as independent and thus the aggregate variance is simply the sum of their individual component's variances. In fact, because of this additive property, the statistical variance was selected to be the measure of imprecision for the estimators [Epitropou et al., 2012].

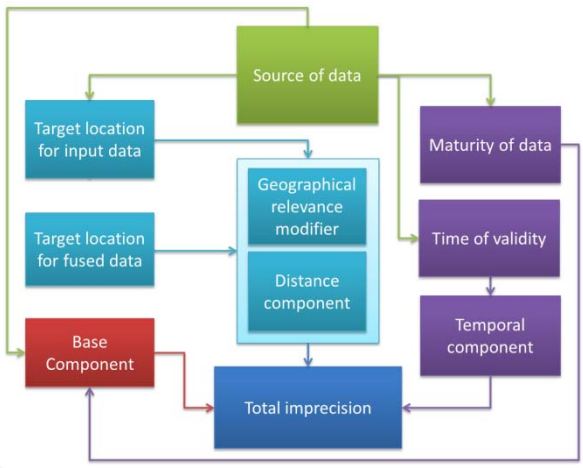


Figure 1: Schematic diagram of the total imprecision model used in the fusion service.

Most input data for fusion is supplied by the DRS and by AirMerge in numerical form, covering both the case of stations and CW forecasts. When performing the above tasks, the Fusion Service performs an estimation of the total imprecision (uncertainty) associated with the numerical data sources used, by using mathematical total imprecision models such as the ones described in [Potempski et al, 2008]. As an extension to those models, the Fusion Service also makes use of land use profile maps (e.g. urban traffic, suburban traffic, residential industrial, rural, etc.) in order to weigh the contributions from different data sources differently based on the characteristics of the terrain, and further parameterize its computation models. Such maps are however

available only for select geographical areas (Helsinki and other locations in Finland), and are statically integrated into the Fusion Service.

3 THE AIRMERGE SUBSYSTEM

3.1 Overview

An important part of AQ information (in particular, Chemical Weather forecasts) is published only in the form of colour-mapped georeferenced images as explained in Epitropou et al. [2010] and Balk et al. [2011]. These are impossible to parse via usual text-mining and screen-scraping techniques used in web mashup-like services. It is important to provide PESCaDO with a specialized service that allows accessing and using CW forecast images as another source of data to use during the Orchestration and Fusion phases. Such a system has already been developed and described in Epitropou et al. [2011] and Balk et al. [2011] in the context of the European Open-access Chemical Weather Forecasting Portal: that of making heterogeneous chemical weather data accessible and mutually comparable, and is currently undergoing linkage and testing within the PESCaDO's infrastructure. These images commonly have geographical spatial resolutions ranging from 1x1 km to 20x20 km, and temporal resolutions from a minimum of one hour to an entire day (Kukkonen et al. [2012]). The reported values usually are maximum or average air pollution concentration values for the selected integration time.

A typical set of such CW models and the resulting images can be found in the European Open-access Chemical Weather Forecasting Portal described by Balk et al. [2011], that has been developed in the frame of COST Action ES0602 (www.chemicalweather.eu). AirMerge is able to convert such image-based concentration maps into numerical, geographically referenced data, accounting for geographical projections, missing data, noise and the differences in publishing formats between different model providers. The result is effectively converting image data back into numerical data, which is now made directly available for a number of applications.

3.2 AirMerge Remote API sample query

AirMerge's data base and services can be accessed through the AirMerge Remote API, which allows exploring the contents of its database, including its internal structure and relational organization using a system of resources exposed through RESTful URLs. This API is also used for interfacing with the PESCaDO system in two ways: i) as a special data node harvested by the DRS and ii) as a direct data source for the Fusion service, meaning that queries directed to the Fusion service automatically cause it to request data from AirMerge directly, without the intervention of the DRS. In addition to simple data retrieval, AirMerge can also perform some numerical processing such as multi-model ensembling (described in Potempski and Galmarini [2008]) through its Remote API, a capability which is also used by the Fusion service.

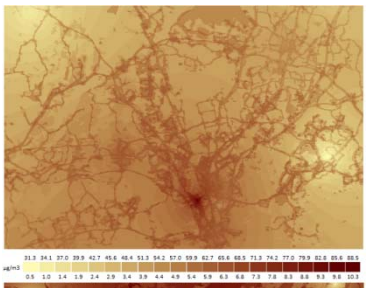
4 RESULTS AND DISCUSSIONS

In Figures 2a and 2b a geographical example of fusion and the resulting fused data imprecision map is presented for NO_2 concentration in the greater Helsinki area on January 1st, 2011. This example fusion corresponds to a typical PEScaDO user query where the user has queried current air quality condition near Helsinki. NO_2 measurements from 11 stations were used as retrieved input data for NO_2.

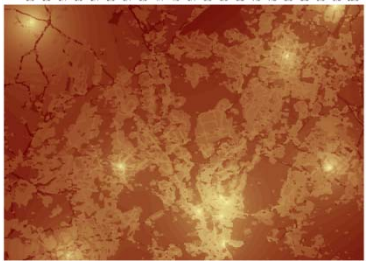
The center of Helsinki is well covered with measurement stations which can be seen from the estimated standard deviation in Figure 2b. In the city centre, the

fused value is estimated to have a standard deviation of only 0.5 μg/m^3 (close to the measurement sites) while at the northern-most parts of the fusion area the standard deviation is estimated to be as much as 10 μg/m3. It is also possible to monitor the relative weights for any individual source of data in the fusion process. Such information about the fused value imprecision and relative source weighting can be readily displayed to the user. In this example the relative weighting for a station in Espoo Leppävaara was visualized in Figure 3. The station in question has been profiled as 'Urban Traffic' and is thus emphasized in the near-by urban road network. However, even if the station in question would have relevance to the urban areas in the city centre for instance, the measurements of Leppävaara station are practically insignificant there according to the fusion service. The reason for this is that the distance causes the total imprecision to increase rapidly; the close-by urban measurement stations in the centre are associated with much lower variance and thus, have larger relative weights associated.

a)



b)



Figures 2a-b: Fused NO_2 concentration map (a) and a standard deviation map for the results (b) in greater Helsinki area in 1st of January 2011 at 9.00 a.m.

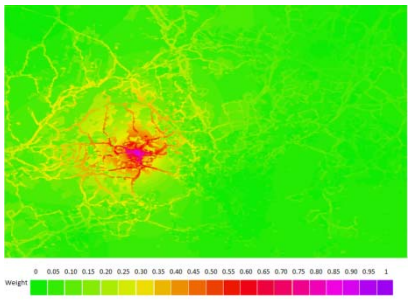


Figure 3: The relative weight assigned for the NO2 measurement data from Leppävaara-station (Suburban traffic) in Espoo for the fusion example in Figure 2a**.**

5 CONCLUSIONS

In this paper, an overview of the PESCaDO service for atmospheric data has been presented and discussed, with a focus on data fusion and the AirMerge system. The immediate benefit of bringing those two systems together will be granting access to an extensive CW database to PESCaDO, currently unobtainable by other means. The bridging requirements between the two platforms have stimulated development of AirMerge's Remote API and the inclusion of new functionalities that will better match PESCaDO's orchestration and Fusion services' needs. This API will also increase the usefulness of AirMerge itself, and its availability to the CW user's community (general public, environmental organizations and authorities, decision makers and scientific community).

6 REFERENCES

Balk, T., Kukkonen J., Karatzas, K., Bassoukos, A., and Epitropou, V., European Open Access Chemical Weather Forecasting Portal, Atmospheric Environment, 38(45), 6917–6922, 2011.

Bergman, K. M., The Deep Web: surfacing hidden value, The Journal Of Electronic Publishing (7), Issue 1, August 2001, doi:10.3998/3336451.0007.104, 2001.

Epitropou, V., Karatzas, K. and Bassoukos, A., A method for the inverse reconstruction of environmental data applicable at the Chemical Weather portal, *Geospatial Crossroads @ GI_Forum '10*, 58–69, Wichmann Verlag, Berlin: ISBN 978-87907-496-9, 2010.

Epitropou, V., Karatzas, K., Bassoukos, A., Kukkonen, J. and Balk, T., A new environmental image processing method for chemical weather forecasts in Europe, *Proceedings of the 5th International Symposium on Information Technologies in Environmental Engineering. Poznan: Springer Series: Environmental Science and Engineering*, 781–791, 2011.

Epitropou, V., Karatzas, K., Kukkonen, J., and Vira, J., Evaluation of the accuracy of an inverse image-based reconstruction method for chemical weather data, *International Journal of Artificial Intelligence*, 9, to appear Autumn 2012.

Ganesh S., Jayaraj, M., Kalyan, V., SrinivasaMurthy, and Aghila S., Ontology-based Web Crawler, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), Vol. 2. IEEE Computer Society, Washington, DC, USA, 337-341, 2004.

Judah, K.; Dietterich, T.; Fern, A.; Irvine, J.; Slater, M.; Tadepalli, P.; Gervasio, M.; Ellwood, C.; Jarrold, B.; Brdiczka, O.; Blythe, J. User initiated learning for adaptive interfaces. IJCAI Workshop on Intelligence and Interaction, 2009 July 13; Pasadena, CA, USA, 2009.

Karatzas, K., A quality-of-urban-life ontology for human-centric, environmental information services, *C21: Towntology, WG1: Ontologies and Information Systems Brussels*, 12–13, 2005.

Karatzas, K. and Kukkonen, J., COST Action ES0602: Quality of life information services towards a sustainable society for the atmospheric environment, ISBN: 978-960-6706-20-2, Thessaloniki: Sofia Publishers, 2009.

Klein Th., Kukkonen J., Dahl Å., Bossioli E., Baklanov A., Fahre Vik A., Agnew P., Karatzas, K., and Sofiev, M., Interactions of physical, chemical and biological weather calling for an integrated assessment, forecasting and communication of air quality, *AMBIO*, 2012, accepted.

Kukkonen, J., Olsson, T., Schultz, D.M., Baklanov, A., Klein, T., Miranda, A. I., Monteiro, A., Hirtl, M., Tarvainen, V., Boy, M., Peuch, V.-H., Poupkou, A., Kioutsioukis, I., Finardi, S., Sofiev, M., Sokhi, R., Lehtinen, K. E. J., Karatzas, K., San José, R., Astitha, M., Kallos, G., Schaap, M., Reimer, E., Jakobs, H., and Eben, K., A review of operational, regional-scale, chemical weather forecasting models in Europe, Atmos. Chem. Phys (12), 1-87, doi:10.5194/acp-12-1-2012, 2012.

Potempski, S. and Galmarini, S., *Est modus in rebus*: analytical properties of multi-model ensembles, *Atmos. Chem. Phys*., 9, 9471–9489, doi:10.5194/acp-9-9471-2009, 2009.

Rodriguez, A., RESTful Web services: The basics, available on line via https://www.ibm.com/developerworks/webservices/library/ws-restful/, 2008 (last accessed 16 Feb. 2012)

Wanner, L., Bosch, H., Bouayad-Agha, N., Bügel, U., Casamayor, G., Ertl, T., Karppinen, A., Kompatsiaris, I., Koskentalo, T., Mille, S., Mossgraber, J., Moumtzidou, A., Myllynen, M., Pianta, E., Rospocher, M., Saggion, H., Serafini, L., Tarvainen, V., Tonelli, S., Usländer, T., and Vrochidis, S., Service-Based Infrastructurs for User-Oriented Environmental Information Delivery, Proceedings of the Enviroinfo Workshop on Environmental Information Systems and Services – Infrastructure and Platforms. Bonn, Germany (http://ceur-ws.org/Vol-679/), 2010

Wanner, L., Vrochidis, S., Tonelli, S., Mossgraber, J., Bosch, H., Karppinen, A., Myllynen, M., Rospocher, M., Bouayad-Agha, N., Bügel, U., Casamayo,r G., Ertl, T., Kompatsiaris, I., Koskentalo, T., Mille, S., Moumtzidou, A., Pianta, E., Saggion, H., Serafini, L., and Tarvainen, V,. Building an Environmental Information System for Personalized Content Delivery. In (Hrebícek J., Schimak G., Denzer R. eds.): *Environmental Software Systems. Frameworks of eEnvironment* - 9th IFIP WG 5.11 International Symposium, Proceedings. IFIP Publications 359, Springer, ISBN 978-3-642-22284-9, pp. 169-176, 2011.