

Service-Based Infrastructure for User-Oriented Environmental Information Delivery

Leo Wanner^{1,2}, Harald Bosch³, Nadjat Bouayad-Agha², Ulrich Bügel⁴, Gerard Casamayor², Thomas Ertl³, Ari Karppinen⁵, Ioannis Kompatsiaris⁶, Tarja Koskentalo⁷, Simon Mille², Jürgen Moßgraber⁴, Anastasia Moutzidou⁶, Maria Myllynen⁷, Emanuele Pianta⁸, Marco Rospocher⁸, Horacio Saggion², Luciano Serafini⁸, Virpi Tarvainen⁵, Sara Tonelli⁸, Thomas Usländer⁴, Stefanos Vrochidis⁶

¹Catalan Institute for Research and Advanced Studies, ²Dept. of Information and Communication Technologies, Pompeu Fabra University, ³Visualization Institute, University of Stuttgart, ⁴Fraunhofer Institute for Optonics, System Technologies and Image Exploitation, ⁵Finnish Meteorological Institute, ⁶Informatics and Telematics Institute, Centre for Research and Technology Hellas, ⁷Helsinki Region Environmental Services Authority, ⁸Fondazione Bruno Kessler
pescado@upf.edu; <http://www.pescado-project.eu>

Abstract. Citizens are increasingly aware of the influence of environmental and meteorological conditions on the quality of their life. The consequence of this awareness is the demand for personalized environmental information, i.e., information that is tailored to their specific context and background. The EU-funded project PESCaDO addresses this demand in its full complexity. It aims to develop a service that supports the user in questions related to environmental conditions in that it searches for reliable data in the web, processes these data to deduce the relevant information and communicates this information to the user in the language of their preference. In this paper, we describe the requirements and the working service-based realization of the infrastructure of the service.

Keywords: environmental information service, personalization, infrastructure, decision support

1 Introduction

Citizens are increasingly aware of the influence of environmental and meteorological conditions on the quality of their life. One of the consequences of this awareness is the demand for high quality environmental information that is tailored to one's specific context and background, i.e., which is personalized. Personalized environmental information may need to cover a variety of themes (such as meteorology, air quality, pollen, and traffic) and take into account a number of specific personal features (health, age, etc.) of the addressee, as well as the intended use of the information. So far, only a few proposals have been made how this information can be facilitated in technical terms. All of these proposals

focus on one theme and only very few of them address the problem of information personalization [Peinel et al., 2000, Karatzas, 2007, Wanner et al., 2010]. PESCaDO (*Personalized Environmental Service Configuration and Delivery Orchestration*) addresses the above task in its full complexity. It takes advantage of the fact that nowadays, the World Wide Web already hosts a great range of services that address each of the above themes, such that, in principle, the required basic data are available. The challenge thus consists, on the one hand, in the discovery of these services and their orchestration, and, on the other hand, in the processing of the obtained data in accordance with the needs of the addressee and the delivery of the gained information in the mode of preference of the addressee. This challenge requires the involvement of an elevated number of rather heterogeneous applications and thus an infrastructure that is flexible and stable enough to support a potentially distributed architecture. In what follows, we first outline the requirements towards the infrastructure of a platform such as PESCaDO, which attempts to integrate all these applications. Then, we present the working infrastructure that has been designed to meet these requirements.

2 The requirements towards PESCaDO’s infrastructure

The requirements towards the infrastructure obviously depend on the tasks that are to be addressed. In the case of PESCaDO, the principal tasks are:

1. Discovery of the environmental service nodes in the web: As already pointed out above, the web hosts a large amount of environmental (meteorological, air quality, traffic, pollen, etc.) services, which include both the numerous (static or dynamic) public webpages that offer environmental information worldwide, as well as any dedicated environmental web services with free access. Especially, the number of meteorological services that cover each major location is impressive. In order to be able to offer citizens targeted information, these services must be exploited, which means that these services must be searched for and indexed such that their data can be accessed when needed.

2. Distillation of the data from webpages: The vast majority of the environmental services offer their data via publicly accessible webpages rather than via web services. To access these data, webpage parsing, information extraction, and text mining techniques are needed. Although these techniques can be tuned to the idiosyncratic ways of presenting environmental (i.e., air quality, meteorological, traffic, ...) data and information, the task of webpage *scraping* remains a very challenging task.

3. Orchestration of the environmental service nodes: Environmental service nodes encountered in the web may require data provided by other service nodes as input data. In order to obtain all necessary data, the environmental service nodes must thus be “orchestrated”, i.e., selected and chained. This presupposes the selection of appropriate protocols and the use of appropriate data interchange formats. To decide which nodes are to be selected over which other nodes, or which nodes fit best together, such criteria as quality of the individual nodes measured by data uncertainty and service confidence metrics derived using machine learning and visual analytics techniques must be taken into account.

4. Fusion of environmental data: Environmental service nodes may provide competing or complementary data on the same or related theme for the same or the neighbour location. To ensure the availability of a most reliable and comprehensive data set as basis for further processing stages, the data from these nodes must be fused. As already in the case of node orchestration, this implies an assessment of the quality of the contributing services and data.

5. Assessment of the data with respect to the needs of the addressee: Once the raw data are obtained, they need to be evaluated and reasoned about in order to infer how they affect the addressee, given his/her personal health and life circumstances and the purpose of the request of the information. Thus, a citizen may request information because he needs it to decide upon a planned action, because he wants to be aware of extreme episodes or because he monitors the environmental conditions in a location. The assessment task obviously presupposes the existence of sufficiently comprehensive domain-specific ontologies and a knowledge base.

6. Selection of user-relevant content and its delivery to the addressee: Not all content deduced from the data by inferences and reasoning is apt to be communicated to the addressee: some of this content would sound trivial or irrelevant. Intelligent content selection strategies that take into account the background of the addressee and the intended use of the information are thus needed to decide which elements of the content are worth and meaningful to be communicated. To deliver the selected content, techniques are required that present the content in a suitable mode (text, graphic and/or table) in the language of the preference of the addressee.

7. Interaction with the user: One should not forget the interaction of the system with the user. The user must be able to formulate his problem in a simple and intuitive format—be it based on natural language or on graphical building blocks. He should equally be delivered the generated information in a suitable form and, as already mentioned above, in the language of his preference.

We are aware of the complexity of each of these tasks. However, given the expertise and the experience of the partners of the PESCaDO Consortium in the corresponding research areas, we are confident to be able to offer an operational PESCaDO service at the end of the lifetime of the project.

3 The PESCaDO infrastructure

In order to meet the above requirements, PESCaDO has opted for a service-based architecture. This architecture is based on a methodology which has been developed for the definition of an open architecture for risk management as provided in the EU FP6 IP project ORCHESTRA [Usländer, 2007] and which has been extended in the FP6 IP project SANY [SANY, 2009] to cover the domain of sensor networks and standard-based sensor web enablement. The focus of this methodology is on a platform neutral specification. In other words, it aims to provide the basic concepts and their interrelationships (conceptual models) as abstract specifications. The design is guided by the methodology developed in the

ISO/IEC Reference Model for Open Distributed Processing (RM-ODP), which explicitly foresees an engineering step that maps solution types, such as information models, services and interfaces specified in information and service viewpoints, respectively, to distributed system technologies. This section illustrates the outcome of this engineering step for the service viewpoint in PESCaDO. Application specific major tasks and actions have been defined as abstract service specifications and can be implemented as service instances on a specific platform. Web service instances for these services are currently being developed. They can be redefined and substituted as needed in the course of the project.

Figure 1 displays a sample, somewhat simplified, workflow with the major application services in action. Two services are not cited in Figure 1 since they are consulted by nearly all other services: the *Knowledge Base Access Service* and the *User Profile Management Service*. Furthermore, the figure does not include the services related to data and information distillation from webpages.

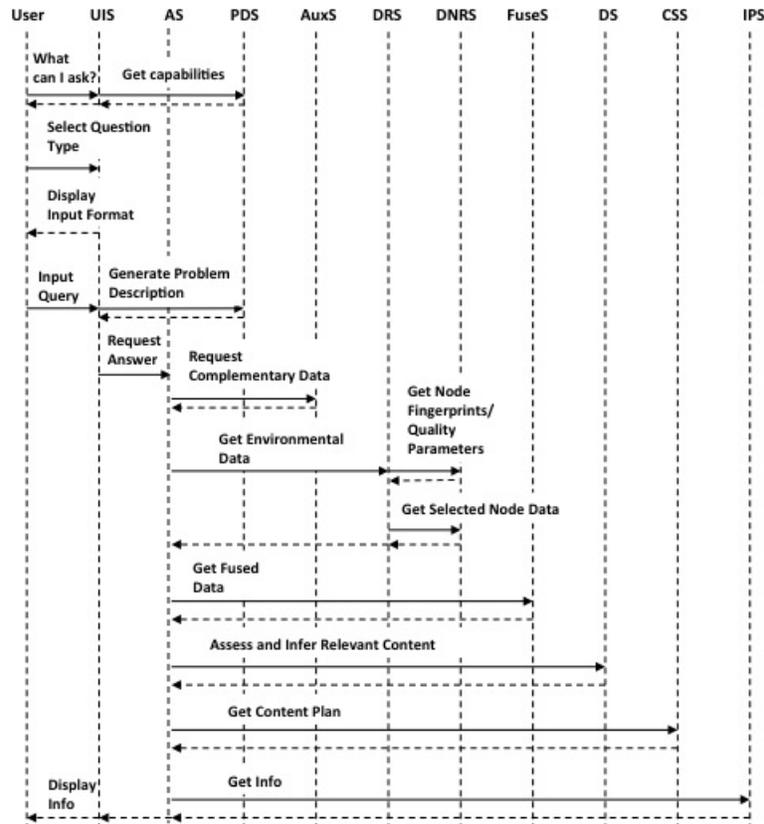


Fig. 1. Sequence diagram for the execution of the services in PESCaDO

A main dispatcher service (called *Answer Service*, AS) controls the workflow and the execution of the services. The user interacts with PESCADO via the *User Interaction Service* (UIS). If unsure about the types of information he can ask for, the user can inquire this information by requesting it from the *Problem Description Service* (PDS).

To ensure a full comprehension of the problem or question of the user, PESCADO decided to operate with controlled graphical and natural language input formats. Once the user has decided what kind of question he wants to submit to the system, the UIS provides the user the corresponding formats. Thereupon, the user can formulate his query, which is translated by the PDS into a formal ontology-based representation understood by the system. Once this is done, the problem description is passed by the UIS to the AS as a ‘Request Answer’ inquiry. Then, the AS assesses what kinds of data beyond environmental data are needed to answer the query of the user and solicit these data from the *Auxiliary Services* (AuxS). For instance, if the user’s query concerns the environmental conditions for a bicycle tour from A to B, the route from A to B must be calculated by a Route Calculation Service.

With the complementary data at hand, the AS can request from the *Data Retrieval Service* (DRS) the environmental data needed to answer the user query. The DRS solicits these data from the environmental nodes that identified by the *Data Node Retrieval Service* (DNRS) as relevant to the user’s query and the complementary data. To speed up retrieval, an off-line indexing is performed. During the indexing procedure, a domain specific search engine accesses the web, discovers, and indexes the environmental service nodes in a local repository. In addition, the retrieved webpages are processed in order to extract environmentally relevant information (e.g. location, environmental measurements, etc.) with the aid of document parsing, web scraping and content distillery techniques, so that each service can be indexed according to this information.

As already pointed out in Section 2, the retrieved nodes may deliver complementary or competing data of varying quality.¹ The *Fusion Service* (FS) applies uncertainty metrics to obtain the optimal and maximally complete data set, which is passed by the AS to the *Decision Service* (DS). The DS converts the data set into knowledge, or content, in that it relates it to the knowledge in PESCADO’s knowledge base, reasons about it, and assesses it from the perspective of its relevance to the user. From this content, the *Content Selection Service* (CSS) compiles a *content plan*, which contains the knowledge to be communicated to the user as answer. The *Information Production Service* (IPS) takes the content plan as input and generates information in the language and mode (text, table, or graphic) of the preference of the user, which then is passed to the user.

¹ For simplicity, we dispense with the illustration of the chaining of service nodes.

4 PESCaDO as part of ICT for Environmental Services

The service-based infrastructure as illustrated above in a sample workflow allows for a maximally flexible realization of the PESCaDO platform: each service (and thus each module of the platform) can be implemented nearly entirely independent from the other services and be run either on a separate machine or on the same machine as the other services. As a matter of fact, many of the services could be used as plug-in modules by other environmental application platforms. How this can be achieved best, needs to be discussed. In any case, a standardization of the communication protocols across the initiatives seems highly desirable.

5 PESCaDO and its consortium

Running from January 2010 to December 2012, PESCaDO is partially funded by the European Commission in its 7th Framework Programme under the contract number ICT-259486. PESCaDO's consortium consists of seven partners: 1. Pompeu Fabra University (UPF), 2. Fraunhofer Institute for Optronics, System Technologies and Image Exploitation (IOSB), 3. Finnish Meteorological Institute (FMI), 4. University of Stuttgart (USTUTT), 5. Foundation Bruno Kessler (FBK), 6. Centre for Research and Technology Hellas (ITI-CERTH), and 7. Helsinki Region Environmental Services Authority (HSY).

The information technologies aspects of the project are covered by CERTH (web-based search), FBK (semantic representation, reasoning strategies, content distillation), UPF (multilingual information generation and human-computer interaction), and USTUTT (visualization and human-computer interaction). The architectural and infrastructure issues are addressed by IOSB. Problems related to uncertainty and confidence metrics of environmental data and information are dealt with by FMI, which, together with HSY also provides its environmental expertise and assumes the validation of the outcome of PESCaDO.

References

- [Usländer, 2007] Usländer, T. (ed.), 2007. Reference Model for the ORCHES-TRA Architecture Version 2.1. OGC Best Practices Document 07-097, http://portal.opengeospatial.org/files/?artifact_id=23286.
- [Karatzas, 2007] Karatzas, K. State-of-the-art in the dissemination of AQ information to the general public. Proceedings of EnviroInfo, Vol. 2. 41–47. Warsaw, 2007.
- [Peinel et al., 2000] Peinel, G., T. Rose and R. San José. 2000. Customized Information Services for Environmental Awareness in Urban Areas. Proceedings of the 7th World Congress on Intelligent Transport Systems. Turin, 2000.
- [SANY, 2009] SANY SensorSA (Sensor Service Architecture of the project SANY). Public OGC Discussion Paper OGC 09-132r1. http://portal.opengeospatial.org/files/?artifact_id=35888&version=1.
- [Wanner et al., 2010] Wanner, L., B. Bohnet, N. Bouayad-Agha, F. Lareau and D. Nicklaß: MARQUIS: Generation of User-Tailored Multilingual Air Quality Bulletins. *Applied Artificial Intelligence*, 24(10), 2010.