Expert Systems

Article

DOI: 10.1111/exsy.12100

Getting the environmental information across: from the Web to the user

Leo Wanner,[1,2] Harald Bosch,[3] Nadjet Bouayad-Agha,[2] Gerard Casamayor,[2] Thomas Ertl,[3] Désirée Hilbring,[4] Lasse Johansson,[5] Kostas Karatzas,[6] Ari Karppinen,[5] Ioannis Kompatsiaris,[7] Tarja Koskentalo,[8] Simon Mille,[2] Jürgen Moßgraber,[4] Anastasia Moumtzidou,[7] Maria Myllynen,[8] Emanuele Pianta,[9] Marco Rospocher,[9] Luciano Serafini,[9] Virpi Tarvainen,[5] Sara Tonelli[9] and Stefanos Vrochidis[7]

(1) Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain Email: leo.wanner@upf.edu
(2) Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain
(3) Institute for Visualization and Interactive Systems, University of Stuttgart, Stuttgart, Germany
(4) Fraunhofer Gesellschaft, Karlsruhe, Germany
(5) Finnish Meteorological Institute, Helsinki, Finland
(6) Aristotle University of Thessaloniki, Thessaloniki, Greece
(7) Center for Research and Technology Hellas, Thessaloniki, Greece
(8) Helsinki Region Environmental Services Authority, Helsinki, Finland
(9) Fondazione Bruno Kessler, Trento, Italy

Abstract: *Environmental and meteorological conditions are of utmost importance for the population, as they are strongly related to the quality of life. Citizens are increasingly aware of this importance. This awareness results in an increasing demand for environmental information tailored to their specific needs and background. We present an environmental information platform that supports submission of user queries related to environmental conditions and orchestrates results from complementary services to generate personalized suggestions. The system discovers and processes reliable data in the Web in order to convert them into knowledge. At runtime, this information is transferred into an ontology-structured knowledge base, from which then information relevant to the specific user is deduced and communicated in the language of their preference. The platform is demonstrated with real world use cases in the south area of Finland, showing the impact it can have on the quality of everyday life.*

Keywords: environmental data discovery, data extraction, data fusion, ontology, data interpretation, decision support, multimodal environmental information generation

1. Introduction

Environmental conditions strongly influence our daily life, and more and more people are aware of this influence. One of the consequences of this awareness is the increasing demand for publicly accessible high quality environmental information. Thus, information on environmental conditions is among the information in the Web that is most often searched for and consulted by the users. According to Google Adwords,[1] there are about 150 million monthly searches with the keyword 'weather', about 14 million searches with the keyword 'environment', and 5 million searches with the keyword 'pollution'. Especially sites that feature meteorological information enjoy a wide-spread popularity because weather greatly influences the leisure and professional activities of a significant share of the population. Biological (pollen) and chemical air quality (AQ) weather sites are equally of increasing demand, although their importance is perceived less by individuals who do not suffer from health conditions aggravated by unfavourable pollen or air pollutant concentrations.

The demand for environmental information by the population led to an explosion of environmental information providers in the Internet. Among them are official institutions such as public and private meteorological institutes,

[1] www.adwords.google.com; consulted in December 2013.

periodicals of broad and local distribution and private companies that make business with advertisement spots they place in the pages along the environmental information. Not all of these providers use measurement networks of the same size and distribution and the same forecasting models, such that it is not surprising that the quality of the information is similarly diverging as the nature of the providers. For instance, as response to the query 'Weather Barcelona' on 3 December 2013, the first hit forecasted a wind speed of 16 km/h and the second 43 km/h; on 20 July 2013, the first claimed the humidity to reach next day 65%, the fifth 93% and the tenth 52%; and on 5 November 2012, the first saw the temperature for the same day to be at 12°C min/18°C max, the second at 10°C min/20°C max and the third at 12°C min/17°C max. Air pollutant information is less frequent in the Web. Also, it is often summarized into the AQ index, such that there is less room for deviance. But, still, diverging concentrations and assessments continue to be rather common.

The user is often lost in the light of these discrepancies or even contradictions. But this is not the only problem: the user is also faced with the problem of the interpretation of the environmental data. Thus, while all of us can judge whether −10°C is cold or not, not all will know how to interpret 179 μ/m^3 of ozone or 52 mg/m^3 of CO concentration, or to judge whether for someone allergic to birch pollen there is a risk of suffering symptoms when the pollen count is 92 $grains/m^3$ and why the wind speed and direction are mentioned at all.

The two central questions for which the user needs an answer in the context of environmental information are, thus, the following:

(1) Given the multitude of environmental information providers in the Web, and the large discrepancy between the information offered by them, what is the information I can trust (i.e. what information is the accurate one)?
(2) What environmental information is relevant to me?

In order to be able to answer these questions, we need intelligent environmental data assessment and user need-tailored information provision technologies. However, surprisingly, few works address the theoretical challenges and/or the design and development of such technologies, let alone the problem of how to identify and cater the correct and relevant environmental Web-based information to the user.

The vast majority of the environmental information services in the Internet present raw data or indices thereof to the public in terms of tables, distribution curves, pictograms or colour scales. Innate to these presentations is 'the same information for all' philosophy. Some of the services intend to offer all the information that might be of relevance to any of the users; others present the information that is assumed to be relevant to a default user (a healthy citizen, with no major background on environmental information). The first expects the user to be able (and willing) to browse through the entire information and decide which information is relevant to him or her; the latter deprives the user with specific needs of information that is relevant to him or her. Both are not in accordance with the design principles of Human–Computer Interaction (Preece *et al.,* 2002).

The need for more intelligent solutions for provision of environmental information has already been voiced in the research community in the late 1990s and in the 2000s; see, for example, Peinel *et al,* (2000), Johansen *et al,* (2001), Bøhler *et al.* (2002) and Karatzas (2007) and further insisted upon more recently; see, for example, Klein *et al.* (2012). It has been argued that such solutions must

(1) Incorporate intelligent data interpretation needed to analyse the course of the measured pollutant concentrations or meteorological conditions and assess their relevance.
(2) Tailor their information to the needs of the users.
(3) Rely upon the textual mode as the central mode.
(4) Offer the information via all modern communication channels.

Some of the prototypical environmental services that attempt to cover (3) are described in Busemann and Horacek (1997), Coch (1998), Bohnet *et al.* (2001) and Bøhler *et al.* (2002). Wanner *et al.* (2010) present the MARQUIS service, which also addresses (1), (2) and (4). However, all of these services have in common that (a) they receive as input data time series from a single measurement network; (b) if they personalize the information, they draw at best on static user profile typologies. Both features are no longer appropriate. Firstly, data obtained from a single source are likely to be less reliable than data counter checked and complemented by data from additional sources. The Web offers a number of complementary and alternative sources for nearly any given region, which lend themselves for use in automated environmental information services. Secondly, although user profile typologies allow for the provision of information that matches the static needs of a user (as, e.g., be aware of an elevated birch pollen concentration in case of a birch pollen allergy), they cannot guide context-oriented choice of the information. The user must be able to inquire the service about the environmental conditions in a specific context and receive active decision support in the language of their preference when planning an activity, which might be influenced by the environmental conditions. The relevant context features span over the user's age (elderly are more sensitive to extreme weather and air pollution conditions than young people) and health conditions (allergic to a specific pollen, suffering from respiratory or heart insufficiency, etc.), as well as the planned activity in combination with background knowledge on the negative consequences of extreme weather or air pollution conditions, and so on.

PESCaDO attempts to overcome the limitations of the state-of-the-art services and address in depth the challenges (1–4) listed earlier.[2] Users of PESCaDO can submit queries on environmental conditions in a region or solicit recommendations with respect to the appropriateness of planned outdoor activities in view of current or expected environmental conditions and in view of their personal profile. The system (a) searches all potentially relevant data sources in the Web; (b) extracts the data; (c) assesses their quality, selecting the most reliable data and fusing, where reasonable, the data from several sources; (d) reasons about them, the inquiry of the user and his or her personal profile record; (e) selects the relevant content to be communicated to the user; and (f) generates a personalized recommendation in English, Finnish or Swedish – depending on the preference of the user. In other words, it develops further a series of cutting edge artificial intelligence technologies in order to cater optimal and personalized information to the user. To integrate these technologies, a service-oriented architecture model is used, which accounts for the asynchronous nature of some of the tasks and ensures maximal modularity and flexibility.

The remainder of the article is structured as follows. The next section assesses the problem of the delivery of user-tailored environmental information, highlighting the requirements for overcoming these limitations. Section 3 presents the service-based architecture of PESCaDO. The following three sections then group the tasks addressed by PESCaDO into three major areas: Section 4 outlines the tasks related to environmental data acquisition, Section 5 deals with the conversion of data into ontology-codified content and Section 6 discusses the processing of the content for the user. In Section 7, the evaluation of PESCaDO is presented, before Section 8 summarizes the achievements of PESCaDO and draws some conclusions.

2. Delivery of user-tailored environmental information revisited

Advanced environmental information delivery that aims to exploit the wealth of environmental data in the Web has to cope with a number of challenges that originate from the nature of the Web, the way environmental data are presented there, the viability of the data and the necessity to account for the needs of different types of users in different contexts. The following five global challenges are essential:

(1) Discovery of environmental service nodes in the Web and data extraction: As already mentioned earlier, the Web hosts a large amount of environmental (meteorological, AQ, pollen, etc.) distributed services, which include both public web pages that offer environmental data worldwide, as well as dedicated environmental Web services with free access. Every now and then, a new service appears or an existing service disappears. There must be thus a dynamic service discovery procedure that identifies and indexes those services that can be of relevance to the targeted users with respect to both the kind of the data they offer and the region they cover. In order to do this with sufficient reliability, this procedure must be able to parse and extract data from heterogeneous formats in which the data are presented – including text, images and numerical data. In particular, data extraction from environmental *heat map* images is a challenge because of their large variety and the current state of the art (Epitropou *et al.,* 2011). Data extraction must, in fact, account for two tasks: (a) content-oriented indexation of the corresponding web pages for targeted search launched upon a specific inquiry by the user and (b) content assessment and relevance-driven selection procedures. Therefore, it must be accurate and exhaustive.

(2) Orchestration of environmental service nodes: Environmental nodes may provide competing or complementary data on the same aspect for the same or the neighbouring location. To ensure the availability of the most reliable and most comprehensive content, the measured and forecasted data from these nodes must be assessed with respect to their trustworthiness and certainty and selected accordingly (if several nodes offer competing data) or fused (if several nodes offer complementary data). For this purpose, uncertainty (or variance or imprecision) metrics and fusion metrics are needed. Imprecision metrics are a standard quality measurement instrument in environmental applications whenever large time series are processed (Li *et al.,* 2007; Kumar, 2008; Potempski & Galmarini, 2009; Park, 2011). However, so far, they have mainly been applied to judge the error margin of the data provided for a specific time and space by a specific measurement source or forecast model.

For high quality environmental information delivery, however, we need also to be able to apply the metric to data of varying spatial and temporal resolution provided by any source or any forecast model, as long as they are in the catchment area of the inquiry of the user. Furthermore, the metric must allow for phenomenon-specific parametrization and calibrate for a variety of needs. For instance, for black ice forecasts, an error of 2–3°C in temperatures around 0°C may cause significant harm, whereas an error of the same size in summer temperatures has, in general, no relevance. That is, we need fusion metrics that

[2] PESCaDO stands for 'Personalized Environmental Service Configuration and Delivery Orchestration'. Its development was supported by the European Commission within its FP7 Programme under the contract number FP7-ICT-248594. Although PESCaDO does not cover the multitude of communication channels, as does MARQUIS, it uses two different modi to convey environmental information: text and graphics.

assess the complementarity of the available data from different sources for a given time and geographical area (s) in order to deduce from all of them the most accurate estimate for the given area(s) or a neighbouring area for which no data are available. A number of contextual parameters must be taken into account, including, for example, the morphology of the landscape of the area in question (street canyon, field, forest, etc.) and the proximity to reference data sources.

(3) From the data to user-relevant information: The extracted and orchestrated data provide an objective environmental snapshot of a given geographical space at a given time. In order to make sense of them, we need to (a) put them into the context of the background knowledge of the environmental domain and (b) evaluate and reason about them in order to determine what data are of relevance to the user and infer how they affect him or her, given his or her personal health and life circumstances and the purpose of his or her request. For instance, a citizen may request information because she wants to decide upon a planned action, be aware of extreme episodes or monitor the environmental conditions in a location. To make this possible, we must build up an environmental background knowledge base (KB), provide mechanisms for automatic mapping of obtained data to content elements and their integration into the KB, and select and refine fuzzy reasoning techniques because 'crisp' reasoning techniques are not appropriate in the light of the vagueness of the percieval.

(4) Catering the content to the user: Not all content in the KB is apt to be communicated to the targeted addressee: some of it would sound trivial or irrelevant (as, e.g., the fact that ozone is an air pollutant or that if it is raining, the streets are wet), some others may not be appropriate because of the profile of the addressee (as, e.g., that 180 μg/m^3 constitute the so-called information threshold in the case of ozone concentration, when compiling information for an environmental expert). Intelligent content selection (CS) strategies must take into account the background of the user and the intended use of the information to decide which elements of the content are worth and meaningful to be communicated. Strategies presented, for example, in Coch (1998), Busemann and Horacek (1997) or Peinel *et al.* (2000), omit CS in this sense altogether.

Once the content has been selected, techniques are required, which present the content in a suitable mode (text, graphic and/or table) and in the preferred language of the addressee. The textual mode is required to be the central mode. That is, a robust full-fledged multilingual and multimodal generator is called for. So far, robust and, at the same time, flexible, high quality multilingual natural language generation has not been in the focus of attention in natural language processing research. State-of-the-art robust text generators are often sentence-template based because they cope with rather restricted sentence constructions and vocabulary (Sripada *et al.,* 2003; Yu *et al.,* 2007; Portet *et al.,* 2009). In the case of personalized environmental information, this is not feasible because of the variety of the vocabulary and sentence constructions.

(5) Interaction with the user: The user must be able to formulate his or her information request or environment-related problem in a simple and intuitive format and receive the generated information in a suitable form. A menue-guided selection of the location, for which the information is desired, or browsing through the blocks of provided information in a webpage – as is most often the case in current environmental services – is not sufficient because of the potential complexity of the requests and the nature of the information provided. More adequate is the formulation of the requests in a *Problem Description* (or *Definition*) *Language* (PDL). The PDL must be expressive enough to capture the request itself, the profile of the user and the context of the request. Furthermore, the PDL must be projectable without any loss of information onto the KB representations in order to be seamlessly related to the corresponding knowledge elements and thus allow for the derivation of the adequate reaction of the system.

The challenge for the presentation of the generated information is twofold: firstly, to be able to select the appropriate mode (text, graphics or table) for distinct chunks of information, depending on the nature of the information and the profile of the addressee and, secondly, to adopt to the information visualization principles (Munzner, 2012) both for each mode and for the layout of the entire information display.

3. Coping with the challenges: the PESCaDO Architecture

Each of the aforementioned challenges can be considered a system design requirement. From the bird's eye view, the architecture of a system for provision of personalized environmental information from the Web that fulfils these requirements consists of four major modules and two repositories (cf. Figure 1). The modules are (A) the environmental data acquisition module, (B) the data-to-content processing module, (C) the user-oriented content processing module and (D) the user interface. The modules (A) and (C) consist, in their turn, of several submodules, each of them assuming one of the tasks listed in the previous section. The repositories are the data base (DB) in which the data extracted from web pages are stored and the KB that contains all the content (including the data imported from the DB) needed for the provision of personalized environmental information.

From the processual perspective, the architecture can be envisaged in terms of two processing pipelines. In the first pipeline, the environmental nodes potentially relevant to the problem space dealt with by PESCaDO are discovered in the Web, and the data from these nodes are extracted and fed into the PESCaDO DB. In the

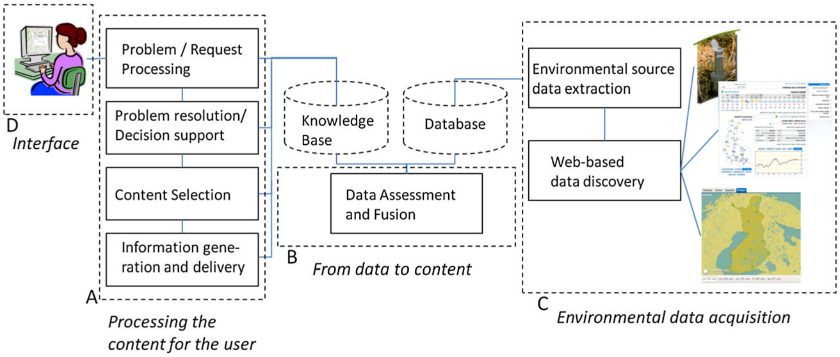

Problem / Request Processing
Problem resolution/ Decision support
Content Selection
Information generation and delivery
A
Processing the content for the user
D
Interface
Knowledge Base
Database
Data Assessment and Fusion
B
From data to content
Environmental source data extraction
Web-based data discovery
C
Environmental data acquisition

Figure 1: PESCaDO's high level architecture.

second pipeline, (a) the request from the user is processed, (b) the request-relevant data in the DB are assessed with respect to their quality/scope and fused where appropriate and integrated into the PESCaDO KB, (c) the problem solution/decision support is reasoned, (d) the content relevant to the user is selected and (e) the corresponding information is generated from this content and delivered to the user.

The two pipelines are asynchronous. Certain tasks in both pipelines are asynchronous as well. Thus, in the first pipeline, data extraction from the already discovered web pages can be carried out while the discovery procedure continues; the same applies in the second pipeline to the reasoning about the static content in the KB and the assessment of the quality of the data/data fusion of the dynamic data introduced depending on the user request. This calls for a distributed architecture that supports asynchronous task management. After an assessment of possible variants, we opted for a service-based architecture, based on the ORCHESTRA/SANY methodology for risk management (Usländer, 2007) and management of sensor (Web) networks (Usländer, 2009). The focus of this methodology is on a platform-neutral specification, which 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, 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. Each independent task and action is thus defined in terms of an abstract service, which is implemented as a platform-specific service. In PESCaDO, the services are realized as Web service instances; they can be redefined and substituted for other applications as required. A main dispatcher service controls the workflow, the execution of the services and their access of the KB. The user interacts with PESCaDO via a separate service that drives the user interface.

In the rest of the section, the two central PESCaDO repositories, the PESCaDO user interface (UI) and the service workflow, are presented, to give the reader an overview of the PESCaDO system. The three process-oriented modules A–C from the preceding text are discussed then subsequently in separate sections.

3.1. *The* PESCaDO *data base*

In order to store, handle and index efficiently the data extracted from the environmental nodes, a DB schema needs to be employed, which is able to store efficiently the clues that comprise the node profiles. The type of data, the need for spatial indexing and generally the nature of the problem let us choose the 52°North Sensor Observation Service,[3] which implements the official *Open Geospatial Consortium* SOS specification 1.0 and depends on PostgreSQL DBMS with the PostGIS extension for representing the data and the spatial information. The key tables of the DB schema, which have been extended to account for the data required for the PESCaDO application,[4] are the following:

(1) *feature_of_interest table* for storing geographical descriptions of regions (cities, municipalities, counties, etc.), for example, *city of Espoo* and *city of Helsinki*, including their geometries.
(2) *Phenomenon table* for representing the types of environmental measurements such as temperature, pollen or chemical pollutant; each phenomenon is identified by its 'phenomenon_id', which corresponds to the uniform resource identifier of the specific named individual found in the PESCaDO ontology (Section 3.2).
(3) *Procedure table* stores metadata of the environmental node that provides the observations.
(4) *Observation table* aggregates the data of an observation event such as, for example, date and time, procedure (sensor or group of sensors), the feature of interest, the phenomenon and the value of the phenomenon.
(5) *Offering table* stores the types of data included in the database and the period for which information is available; an example of a type of offering is AirQualityDataType class of PESCaDO ontology.

[3] https://wiki.52north.org/bin/view/Sensornet/SensorObservationService#SOS_tutorial
[4] The DB extended schema continues to conform to the Observations and Measurements 2.0 (O&M 2.0) standard.

3.2. *The* PESCaDO *knowledge base*

The KB acts as the main data structure that is accessed by all services of the second pipeline of the system.[5] It is realized as an instantiated Web Ontology Language (OWL) ontology. The static class partition of the ontology is built partly manually and partly exploiting automatic key-phrase extraction techniques (Tonelli *et al.,* 2011) and further extended by the available geographical and provenance ontologies *geosparql* [6] and PROV-O.[7] The information relevant to a user request (such as the context of the inquiry and measured data) is dynamically instantiated. See (Moßgraber & Rospocher, 2012) for details on the management of the KB.

The core part of the ontology of the PESCaDO KB comprises three main modules: (a) the *Problem Description Language* module (PESCaDO-PDL), (b) the *Data* module (PESCaDO-Data) and (c) the *Conclusions* module (PESCaDO-Conclusions). That is, in contrast to the common tendency to use ontologies in environmental applications only for data representation (Zagorulko & Zagorulko, 2010) and/or for reasoning (Ceccaroni *et al.,* 2004), in PESCaDO, ontologies form the integrative core of the system.

PESCaDO-PDL formally describes all aspects of decision support requests that the user can submit to the system. It consists of three interrelated submodules: PESCaDO-Request, PESCaDO-User and PESCaDO-Activity. PESCaDO-Request describes a taxonomy of request types supported by the system (e.g., ‘Is there any health issue for me?’ and ‘Do environmental conditions require some administrative actions?’). PESCaDO-User defines the building blocks of the profile of the user involved in the request. Examples of the aspects modelled in this module are the user typology (e.g., ‘end-user’ or ‘administrative user’), the age of the user, the gender, diseases or allergies the user may suffer from, and so on. PESCaDO-Activity describes the activities that the user may want to undertake and that may affect the decision support provided by the system – among them, for example, ‘physical outdoor activity’ and ‘using public transportation’. Figure 2 shows the class hierarchy of the three submodules composing the PESCaDO-PDL.

PESCaDO-Data formally describes the environmental data accessed and manipulated by the system to produce personalized content, among them meteorological (temperature, wind speed, precipitation, etc.), pollen and AQ data (e.g., NO_2, PM10 and AQ index). Environment-related data such as traffic density and road conditions are also represented. All the necessary details needed to comprehensively describe observed, forecasted and historical data are described, including quantitative and qualitative measurement values and the mapping between the two,[8] the period covered by the data and the type of the data (e.g., instantaneous, average, minimum and maximum). Detailed information on the environmental nodes providing the data is also captured, including its type (e.g., measurement station, website and Web service), geographical location and confidence value. The characterization of environmental data and nodes in the PESCaDO ontology is illustrated in Figure 3.

The ontological representation of the data processed by the PESCaDO system is used to integrate the input data coming from heterogeneous sources for decision-making purposes and obtained by different techniques – for instance, by querying environmental Web services or by distilling data from texts and images offered by environmental websites.

PESCaDO-Conclusions formally describes the personalized content produced by the PESCaDO system by processing the problem description and the available data. Typical content includes (a) conclusions such as warnings, recommendations and suggestions that may be triggered by environmental conditions; (b) exceedances of air pollutants limit values that may be detected from the data; and (c) data aggregations and data analysis results. An example of a warning type encoded in the PESCaDO-Conclusions module is presented in Figure 4, together with the associated warning message to be reported to the users (available in all three languages supported in PESCaDO: English, Finnish and Swedish).[9]

Details on the confidence of the system about this content may also be represented by means of a [0…1] weight. Furthermore, PESCaDO-Conclusions contains logico-semantic relations (LSRs) to support the automatic generation of personalized information (Bouayad-Agha *et al.,* 2012a). An LSR is a domain-independent relation that expresses a logical link between domain entities. LSRs support text planning, where they are used to enforce coherence in the text (Section 2). The following LSRs are captured: Implication, Cause, Violation of Expectation, Contrast, List and Elaboration. They are formalized as concepts,[10] such that for each type of LSR an OWL class is defined. Figure 5 shows (a) the hierarchy of LSRs and (b) the formal definition of the Cause relation.

3.3. *The* PESCaDO *user interface*

The PESCaDO UI is designed to support the user in the formulation of meaningful queries/problem descriptions and to adequately present the personalized information offered by the system. The UI is realized as an intelligent wizard dialogue

[5] The KB is available at http://pescado-project.upf.edu/. See Rospocher and Serafini (2012) for a more detailed presentation.

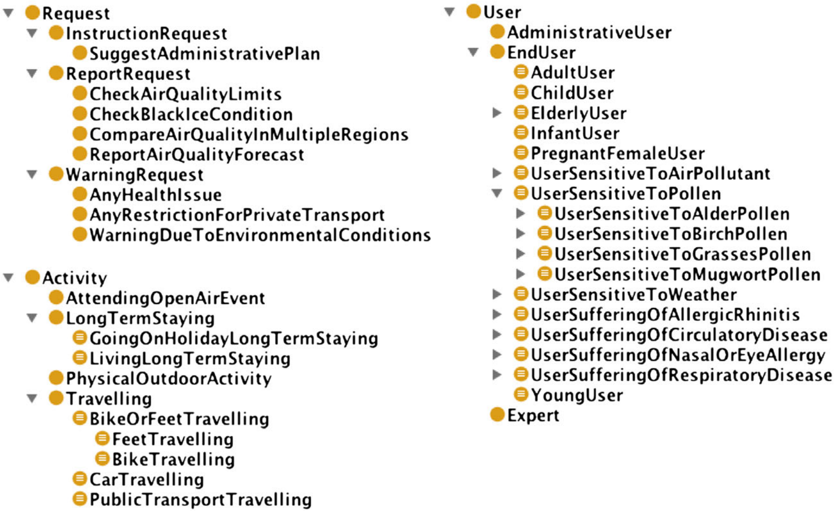
[6] version 1.0 (http://www.opengeospatial.org/standards/geosparql)

[7] Working Draft 03 May 2012 (http://www.w3.org/TR/prov-o/).

[8] For instance, ‘moderate concentration of birch pollen’ corresponds to a concentration between 10 and 100 grains/m^3 of air, ‘mild temperature’ in winter time in Finland corresponds to a temperature up to -10°C and so on.

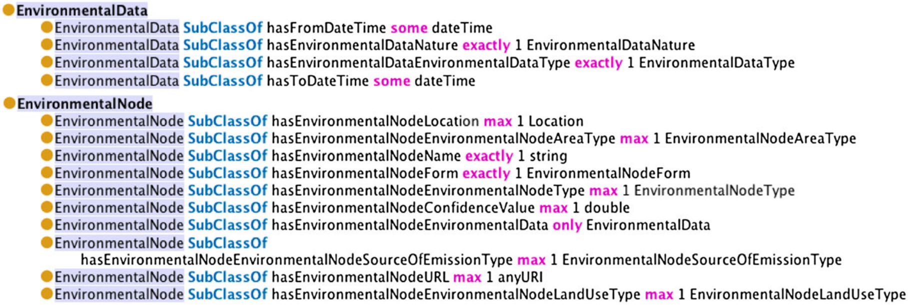
[9] Such warnings are legal texts that cannot be changed and that must thus be displayed to the user as they are.

[10] Logico-semantic relations are defined as concepts rather than as object properties because the arity of some LSRs is higher than 2, while OWL allows us to express only binary properties.



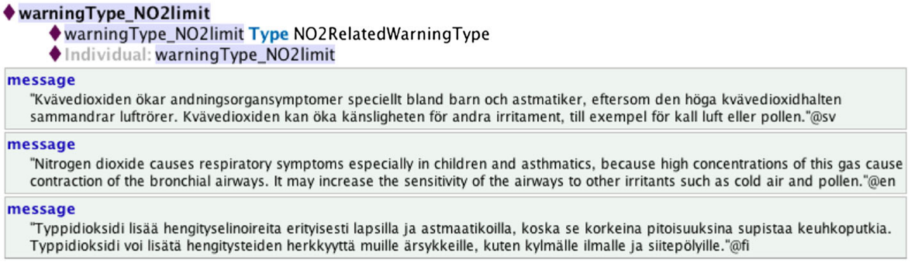
```
▼ Request
    ▼ InstructionRequest
        SuggestAdministrativePlan
    ▼ ReportRequest
        CheckAirQualityLimits
        CheckBlackIceCondition
        CompareAirQualityInMultipleRegions
        ReportAirQualityForecast
    ▼ WarningRequest
        AnyHealthIssue
        AnyRestrictionForPrivateTransport
        WarningDueToEnvironmentalConditions

▼ Activity
    AttendingOpenAirEvent
    ▼ LongTermStaying
        GoingOnHolidayLongTermStaying
        LivingLongTermStaying
    PhysicalOutdoorActivity
    ▼ Travelling
        BikeOrFeetTravelling
            FeetTravelling
            BikeTravelling
        CarTravelling
        PublicTransportTravelling

▼ User
    AdministrativeUser
    ▼ EndUser
        AdultUser
        ChildUser
      ▶ ElderlyUser
        InfantUser
        PregnantFemaleUser
      ▶ UserSensitiveToAirPollutant
      ▼ UserSensitiveToPollen
          ▶ UserSensitiveToAlderPollen
          ▶ UserSensitiveToBirchPollen
          ▶ UserSensitiveToGrassesPollen
          ▶ UserSensitiveToMugwortPollen
      ▶ UserSensitiveToWeather
      ▶ UserSufferingOfAllergicRhinitis
      ▶ UserSufferingOfCirculatoryDisease
      ▶ UserSufferingOfNasalOrEyeAllergy
      ▶ UserSufferingOfRespiratoryDisease
        YoungUser
    Expert
```

Figure 2: *Class hierarchy of the three submodules composing the PESCaDO-PDL.*



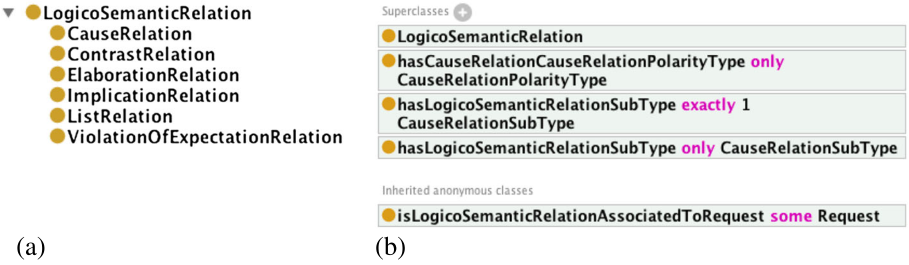
```
EnvironmentalData
    EnvironmentalData SubClassOf hasFromDateTime some dateTime
    EnvironmentalData SubClassOf hasEnvironmentalDataNature exactly 1 EnvironmentalDataNature
    EnvironmentalData SubClassOf hasEnvironmentalDataEnvironmentalDataType exactly 1 EnvironmentalDataType
    EnvironmentalData SubClassOf hasToDateTime some dateTime
EnvironmentalNode
    EnvironmentalNode SubClassOf hasEnvironmentalNodeLocation max 1 Location
    EnvironmentalNode SubClassOf hasEnvironmentalNodeEnvironmentalNodeAreaType max 1 EnvironmentalNodeAreaType
    EnvironmentalNode SubClassOf hasEnvironmentalNodeName exactly 1 string
    EnvironmentalNode SubClassOf hasEnvironmentalNodeForm exactly 1 EnvironmentalNodeForm
    EnvironmentalNode SubClassOf hasEnvironmentalNodeEnvironmentalNodeType max 1 EnvironmentalNodeType
    EnvironmentalNode SubClassOf hasEnvironmentalNodeConfidenceValue max 1 double
    EnvironmentalNode SubClassOf hasEnvironmentalNodeEnvironmentalData only EnvironmentalData
    EnvironmentalNode SubClassOf
        hasEnvironmentalNodeEnvironmentalNodeSourceOfEmissionType max 1 EnvironmentalNodeSourceOfEmissionType
    EnvironmentalNode SubClassOf hasEnvironmentalNodeURL max 1 anyURI
    EnvironmentalNode SubClassOf hasEnvironmentalNodeEnvironmentalNodeLandUseType max 1 EnvironmentalNodeLandUseType
```

Figure 3: *Characterization of environmental data and nodes in the PESCaDO ontology.*



```
♦ warningType_NO2limit
    ♦ warningType_NO2limit Type NO2RelatedWarningType
    ♦ Individual: warningType_NO2limit
message
    "Kvävedioxiden ökar andningsorgansymptomer speciellt bland barn och astmatiker, eftersom den höga kvävedioxidhalten
    sammandrar luftrörer. Kvävedioxiden kan öka känsligheten för andra irritament, till exempel för kall luft eller pollen."@sv
message
    "Nitrogen dioxide causes respiratory symptoms especially in children and asthmatics, because high concentrations of this gas cause
    contraction of the bronchial airways. It may increase the sensitivity of the airways to other irritants such as cold air and pollen."@en
message
    "Typpidioksidi lisää hengityselinoireita erityisesti lapsilla ja astmaatikoilla, koska se korkeina pitoisuuksina supistaa keuhkoputkia.
    Typpidioksidi voi lisätä hengitysteiden herkkyyttä muille ärsykkeille, kuten kylmälle ilmalle ja siitepölyille."@fi
```

Figure 4: *Excerpt of the PESCaDO-Conclusions.*



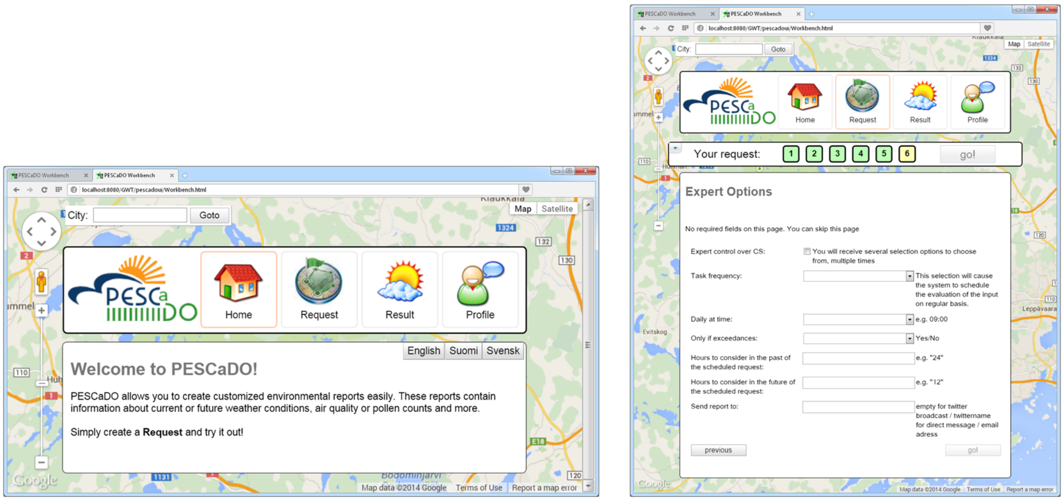
```
▼ LogicoSemanticRelation
    CauseRelation
    ContrastRelation
    ElaborationRelation
    ImplicationRelation
    ListRelation
    ViolationOfExpectationRelation
```

(a)

```
Superclasses
LogicoSemanticRelation
hasCauseRelationCauseRelationPolarityType only
CauseRelationPolarityType
hasLogicoSemanticRelationSubType exactly 1
CauseRelationSubType
hasLogicoSemanticRelationSubType only CauseRelationSubType

Inherited anonymous classes
isLogicoSemanticRelationAssociatedToRequest some Request
```

(b)

Figure 5: *(a) Logico-semantic relation class hierarchy and (b) formal definition of Cause relation.*



Welcome to PESCaDO!

PESCaDO allows you to create customized environmental reports easily. These reports contain information about current or future weather conditions, air quality or pollen counts and more.

Simply create a **Request** and try it out!

(a)

Your request: 1 2 3 4 5 6 go!

Expert Options

No required fields on this page. You can skip this page

Expert control over CS: □ You will receive several selection options to choose from, multiple times

Task frequency: This selection will cause the system to schedule the evaluation of the input on regular basis.

Daily at time: e.g. 09:00

Only if exceedances: Yes/No

Hours to consider in the past of the scheduled request: e.g. "24"

Hours to consider in the future of the scheduled request: e.g. "12"

Send report to: empty for twitter broadcast / twittername for direct message / email adress

previous go!

(b)

Figure 6: *(a) UI wizard; (b) Interactive inquiry composition, page 6.*

hovering over a geographical map and interactive elements that are placed at the corners of the map (cf. Figure 6(a)). The geographic area of interest is the natural link between the user query/problem description generation and the visual result presentation on the same canvas. The wizard allows a free navigation between the pages in order to avoid patronizing the users. The UI validates each user input to identify inconsistent queries and lack of information in the required input fields, based on the already provided information. In the case of inconsistencies, the relations between input fields are highlighted to enable the user to correct his or her input.

The wizard has two levels of navigation. The top level determines the content of the lower part of the wizard and is separated into the tabs 'Home', 'Request', 'Result' and 'Profile'. The Request Tab contains the second level of navigation: the request pages, numbered from 1 to 6. These pages contain the input elements to formulate a complete request grouped into separate pages by their theme and by their interdependency: (1) selecting the request type; (2) selecting the activity for which environmental information related decision support is requested; (3) selecting the geographical context of the request; (4) selecting the temporal context of the request; (5) providing personal information; and (6) advanced functionality such as request scheduling and interactive CS (cf. Figure 6(b)). In all pages, options made invalid by previous input are greyed out.

The interactive formulation of the query is controlled in that (a) highlightings warn the user when invalid information was supplied or mandatory information is missing when the user's cursor is near the 'next' button,and (b) errors are marked with a red line indicating with what other input the information is conflicting. These errors become apparent after the user confirmed his or her input by navigating away from the current page and coming back again later. The 'start' button is only enabled after the query is in a serviceable state, that is, the explicit rule set derived from the requirements of the PDL is satisfied. Otherwise, the problematic areas are marked in the page overview and on the individual pages.

For the validation of the user inquiry, the UI draws upon the PESCaDO-PDL (Section 2) during the initialization stage in that it derives a set of rules from PDL's subclass relations, class restrictions and generic properties that model inquiries it can handle (cf. Table 1) for illustration. During the user session, these rules are applied to the inserted inquiries.

The personalized information returned by the system as response to the processed inquiry is displayed as integral part of the UI wizard (Section 2) on a designated tab.

3.4. *Services of the processing modules*

As pointed out earlier, each independent task carried out by PESCaDO's modules is implemented in terms of a service. Accordingly, we can distinguish between services that deal with (a) the acquisition of environmental data from the Web, (b) the assessment and fusion of data and their conversion into content, (c) the processing of content in the context of an inquiry of the user and (d) user management. Table 2 summarizes the main PESCaDO services.

Depending on the concrete application, the services can be assembled to a specific workflow that serves best the application. Figure 7 shows the workflow for the provision of personalized decision support information to citizens via the Web once the acquisition of the data from the Web has been accomplished. In the case of the provision of information on the quality of a specific environmental data source, the composition of the workflow would be rather different (for instance, it would not involve most of the content processing services).

4. Environmental data acquisition

The environmental nodes, which consist of Web portals and sites, are highly distributed all over the Web, and each of them offers the data in terms of both texts and images in a proprietary format. This makes the problem of discovery and extraction of environmental Web-based nodes a serious challenge.

Table 1: *Examples of the derivation of rules from the problem description language constructs*

Ontology relation	Resulting rule	Rule effect
Y subClassOf X	$X \wedge \neg (Z_1 \vee \ldots \vee Z_n)$	Requires Y
S hasSomeValuesFrom $\{Z_1 \ldots Z_n\}$	S	Requires L
not(S hasSomeValuesFrom $\{Z_1 \ldots Z_n\}$)	S	Forbids Y
X hasOnlyValuesFrom (Y)	X	Requires Y
X hasExactCardinality (1, Y)	X	Requires Y
X hasMinCardinality (2, Y)	X	Requires count: Y

Z_i are the siblings of Y, and L is the least common ancestor of all Z_i. The prefix count stands for 'number of occurrences of Y'.

Table 2: *The services in* PESCaDO

Data acquisition services	Node discovery service	Searches for the environmental nodes in the Web.
	Data extraction service	Extracts multimodal environmental data from the discovered environmental nodes and stores them in the data repository.
	Data indexing service	Indexes the environmental data to be stored in the data repository.
	Data retrieval service	Facilitates access to the data repository.
Data to content service	Fusion service	Acts as a bridge service between the two blocks of services in that it collects the data related to the inquiry of the user from the DB, assesses, extrapolates and fuses them and feeds them into the ontology.
Content processing services	Answer service	Main dispatcher service that supervises the control and the information flows between all services active during the processing of an inquiry of the user.
	Problem description generation service	Maps the inquiry of the user onto the formal problem description, which can be instantiated in the ontology and processed further.
	Related aspects computation service	Identifies content in the ontology that is indirectly related to the inquiry of the user (e.g., the weather conditions in an area for which the user requested a pollen forecast).
	Knowledge base access service	Handles the instantiation of new knowledge elements in the ontology (i.e., writing into the ontology and reading out of content from the ontology).
	Decision service	User request interpretation and reasoning service; identifies all (explicit and deduced) content in the ontology that is relevant to the inquiry of the user.
	Content selection service	Selects the content (from the content pool identified by the DS) that is to be communicated to the user.
	Mode determination service	Determines the communication mode (text or graphic) for the individual chunks of content selected by the CSS.
	Information production service	Generates textual and graphic information from the selected content.
User services	User profile management service	Serves as a bridge between the user and PESCaDO by maintaining the user profile information.
	User interaction service	Ensures the communication between PESCaDO and end users and between the expert user and interactive services.

4.1. *Environmental node discovery*

PESCaDO's *Node Discovery Service* addresses environmental node discovery as a domain-specific search problem. Therefore, we apply methodologies developed in this field. However, given that we are interested only in web pages that contain observed or forecasted environmental data, excluding, for example, scientific or general public background information on environment with examples of data, the problem constitutes a serious challenge that cannot be solved by a standard domain-specific search engine that would retrieve all pages indiscriminately. To address this challenge, we combine two types of methodologies of domain search: (a) domain-specific query submission to a general-purpose search engine, followed by a post-processing phase and (b) focused (or directed) crawling; see Figure 8 for the architecture of the Node Discovery Service.

In the case of domain-specific query submission to a general-purpose search engine, basic and extended queries are distinguished (Moumtzidou *et al.,* 2012b). The basic queries are compiled by combining environment-related

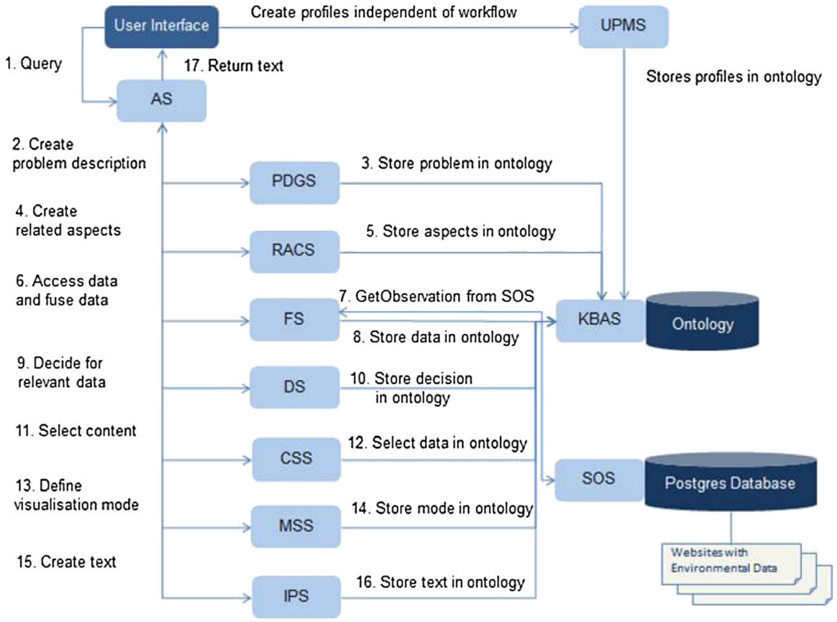

User Interface
Create profiles independent of workflow
UPMS
1. Query
17. Return text
AS
Stores profiles in ontology
2. Create problem description
PDGS
3. Store problem in ontology
4. Create related aspects
RACS
5. Store aspects in ontology
6. Access data and fuse data
FS
7. GetObservation from SOS
8. Store data in ontology
KBAS
Ontology
9. Decide for relevant data
DS
10. Store decision in ontology
11. Select content
CSS
12. Select data in ontology
SOS
Postgres Database
13. Define visualisation mode
MSS
14. Store mode in ontology
Websites with Environmental Data
15. Create text
IPS
16. Store text in ontology

Figure 7: *Service workflow for the provision of personalized decision support information to citizens.*

keywords (e.g., weather, temperature, ozone and AQ) obtained from PESCaDO ontology with geographical data (e.g., city names) retrieved using geographical Web services (among them, e.g., GeoNames) (cf. a sample query):

```
weather + Helsinki.
```

The extended queries are generated by enhancing the basic queries with domain-specific keywords using the *keyword spice* technique (Oyama *et al.*, 2004) (cf. a sample query):[11]

```
weather +forecast -instrument -part -surface
        -comment -climate +Helsinki
```

As the general-purpose search engine application programming interface (API), the Yahoo! BOSS API is used.

In the case of focused crawling, a crawler explores the Web in a directed fashion and collects other nodes that satisfy specific criteria related to the content of the source pages and the link structure of the Web. The PESCaDO crawler is built upon the Apache Nutch crawler, which it extends by integrating text and hypertext classification. More precisely, the crawler retrieves text from specific parts of the web page (i.e. anchor text, text around link and URL itself) and filters the URLs that are visited through machine learning techniques as proposed by Tang *et al.* (2004).

Because the web pages returned by both techniques include a significant number of irrelevant hits, it is essential to introduce a post-processing classification that improves precision. For this purpose, a Support Vector Machine (SVM) classifier is used. We employed the LIBSVM library (Chang & Lin, 2011) with a binary C-Support Vector Classification, using as kernel the radial basis function. The classifier is trained on manually annotated websites and textual features extracted using the KX environment (Pianta & Tonelli, 2010), which identifies a list of concepts in a document and ranks them with respect to their relevance taking into account the language and the structure of the webpage.

Although the discovery procedure is automatic, an administrator can intervene via an interactive graphical UI in order to select geographic regions of interest to perform the discovery and parameterize the post-processing through a relevance feedback-based interactive classification interface (Vrochidis *et al.*, 2012a).

The combination of multiple methodologies in the discovery procedure proved to be highly effective. Thus, experiments have shown that the initial precision of the Yahoo! BOSS API with basic queries reached only 29%. Using extended queries, we already achieve around 70%. The SVM classification of the initial result lists further improves the discovery performance: with basic queries, the precision is increased by 38.5% to 67.5%, while with extended queries, it is increased by 11% to about 81% (Moumtzidou *et al.*, 2012b). High precision is of great importance, given that data extraction techniques are applied to all derived web pages in order to extract environmental data relevant to users of PESCaDO.

4.2. *Extraction of data from environmental nodes*

The discovered environmental web pages contain relevant data as texts and images. The goal of the extraction task is to populate the PESCaDO DB with environmental data by extracting them from the discovered web pages and feed them into the DB. As mentioned in Section 1, we use a SOS server

[11] Keyword spices are Boolean expressions of keywords that are produced by applying supervised machine learning techniques to a set of predefined web pages. In PESCaDO, we use the ID3 algorithm implementation of the WEKA workbench http://www.cs.waikato.ac.nz/ml/weka/, (Witten *et al.*, 2011).

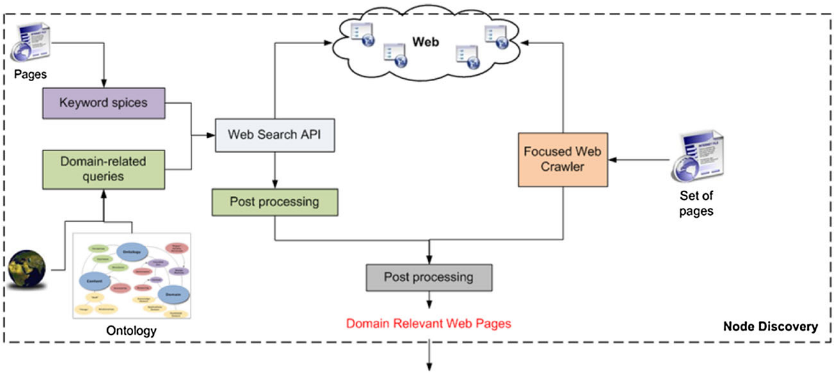


Figure 8: *Architecture of the node discovery service.*

as interface to the DB, which facilitates the indexing and retrieval of real time or archived spatial data produced by all kinds of sensors (including, e.g., sites) – as required in the context of PESCaDO.

4.2.1. Extraction of data from textual sources In textual sources, environmental data are mainly displayed in web pages in a tabular format. To parse the HTML structure of the page and retrieve the relevant data, a number of routines have been implemented that act in sequence. A page is first downloaded and its HTML structure is converted into a tree-like format.[12] Then, the following routine sequence is applied:

(1) Go through the textual content of the page to look for time and date information.
(2) Parse the tree and isolate the sub-trees containing a table.
(3) For each subtree, search the cells containing environmental data and extract them.
(4) Connect the environmental data with the corresponding time/date information and convert them into the format compliant with the PESCaDO ontology.

The extracted data need to be very precise because they are further processed in the PESCaDO service chain to compute exceedances and generate forecasts and recommendations. Given that the formats in which the data are presented in the web pages are very idiosyncratic, a certain specific adaptation to prominent known pages of the regions frequented by the users[13] cannot be avoided, such that all data are extracted and carefully checked by specific scripts. This guarantees that the data extracted from known web pages achieve an accuracy of 100%. In order to also offer a generic data extraction from web pages of any region in Europe, a generic data identification and extraction procedure has been developed. This procedure is based on routines that identify triggering elements in the HTML page, for instance, the Celsius degree symbol '°', 'bar' as measurement unit for pressure, the '%' symbol for humidity and the 'm/s' or 'km/h' abbreviations for wind speed. Upon the identification of a triggering element, the HTML tree is searched, and a fixed number of nodes are explored for numbers (values), dates and names of locations, which are then retrieved using fixed patterns and lists in the form of regular expressions. However, an evaluation of this generic procedure revealed a number of limitations. For instance, the triggering elements are not always present in the page, and the retrieved data may be largely incomplete. In view of this experience, an adaptation of the data extraction procedure is necessary when the service is ported to a new region.

To cope with multilinguality, the extraction routines avoid to search for concrete word patterns (such as 'Last Update' and 'Weather Forecast' in English). Instead, they draw upon the HTML source code: the code is not displayed to the user, and therefore, it remains generally the same, even if the language option of the page is changed by the user. To understand how the environmental data are internally encoded in the page sources and to identify generic source code that can be used as a hint for the presence of specific environmental data, the same pages in different languages were retrieved and carefully studied over a period. After observing how the code changes with different weather conditions, the extraction routines were adapted to retrieve the environmental data exclusively from HTML tags and language-independent elements in the page (e.g., images with the corresponding attributes).

4.2.2. Extraction of data from image sources Images (especially heatmaps) are a very common means for the presentation of AQ and pollen information in a web page. While this presentation format is informative for the casual user, it poses a big challenge for the extraction of concrete values. In order to address this problem, a semi-automated procedure has been developed in PESCaDO. This procedure combines a heatmap annotation tool (AnT) with text and image (mainly heatmap) processing modules (Vrochidis *et al.*, 2012b); see Figure 9.

The AnT facilitates the annotation process for the user. AnT provides two different interaction methods, which are realized on two different data views. The first data view is designed as a simple TreeView, which represents the

[12] For this task, the HTML::TreeBuilder library was used.
[13] In PESCaDO's use cases, this was Finland.

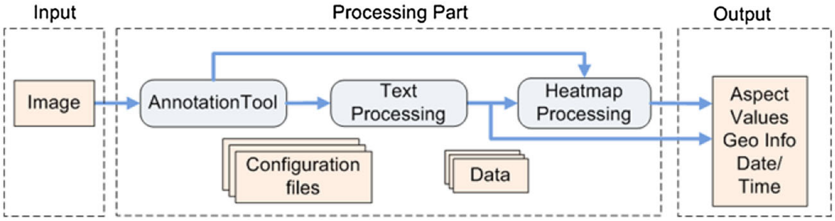


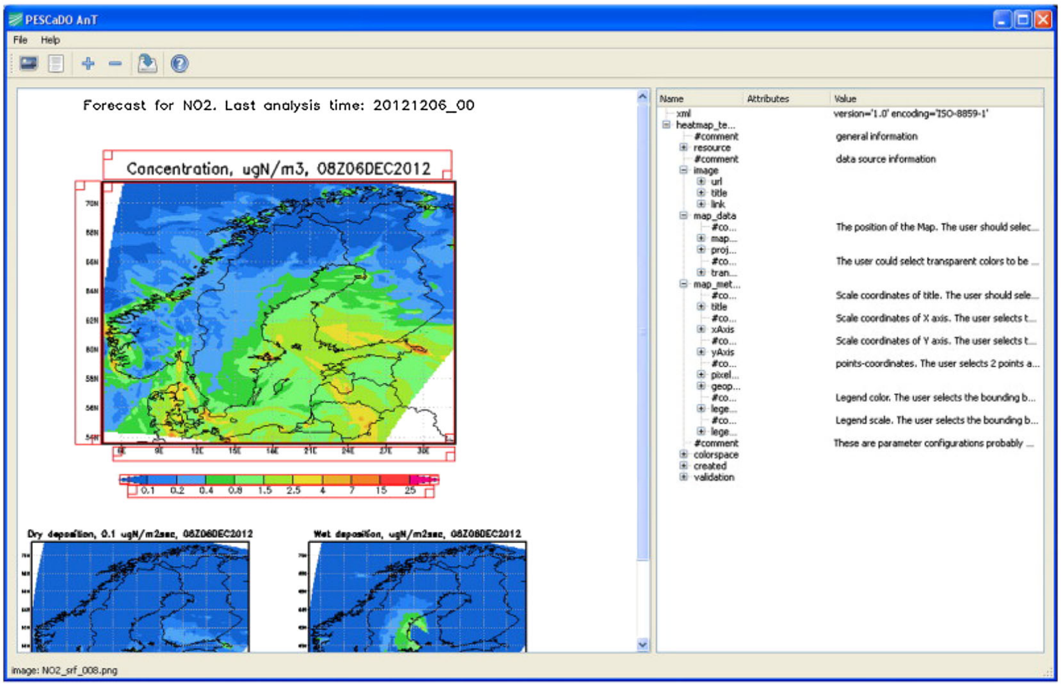
Figure 9: *Extraction of data from image sources in* PESCaDO.

underlying XML data structure and its entries as a traversable tree. The second data view is conceived as a GraphicsView, which is capable of interpreting and viewing the selected datasets graphically. This view is used to draw regions of interest or points of interest as overlays over the heatmap. Figure 10 shows AnT with an already loaded heatmap of a typical site with heatmaps.

The text processing module focuses on retrieving the textual information captured in the image using text extraction and processing techniques through a two-step procedure. The first step includes the application of optical character recognition on the following parts of the initial image: title, colour scale, *x* and *y* axes of the map and searching for potential text strings containing relevant information to the heatmap itself. The second step involves the application of text processing based on heuristic rules in order to correct, extract and understand the semantic information encoded in the aforementioned locations.

The image processing module extracts data from different models and coordinate systems. It is realized by the AirMerge engine (Epitropou *et al.*, 2011), which is a processing framework with the primary purpose of extracting environmental data from heatmaps. The most important component of AirMerge is the AirMerge Core Engine, which performs the conversion of image data (heatmaps) into numerical grid data. The Core Engine performs the extraction of data from heatmaps using a processing chain that consists of two main procedures: (a) the screen scraping procedure, where raw RGB pixel data are extracted from heatmaps, classified according to a colour scale and mapped to ranges of numerical values and (b) the reconstruction of missing values and data gap procedure, which deals with noisy elements on heatmaps.

The annotation tool accepts as input an image containing a heatmap and produces a configuration file that contains information on the heatmap elements. This information is then forwarded to the text processing module, which extracts textual data from the corresponding image. The produced data are sent along with the configuration file to the heatmap processing module, which processes the heatmap located inside the image. The output of this procedure is an XML file in which each geographical coordinate of the initial heatmap is associated with a value.


PESCaDO AnT
File
Help
Forecast for NO2. Last analysis time: 20121206_00
Concentration, ugN/m3, 08Z06DEC2012
0.1
0.2
0.4
0.8
1.5
2.5
4
7
15
25
Dry deposition, 0.1 ugN/m2sec, 08Z06DEC2012
Wet deposition, ugN/m2sec, 08Z06DEC2012
image: NO2_srf_008.png

Name	Attributes	Value
xml		version='1.0' encoding='ISO-8859-1'
heatmap_te...		
#comment		general information
resource		
#comment		data source information
image		
url		
title		
link		
map_data		
#co...		The position of the Map. The user should selec...
map...		
proj...		
#co...		The user could select transparent colors to be ...
tran...		
map_met...		
#co...		Scale coordinates of title. The user should sele...
title		
#co...		Scale coordinates of X axis. The user selects t...
xAxis		
#co...		Scale coordinates of Y axis. The user selects t...
yAxis		
#co...		points-coordinates. The user selects 2 points a...
pixel...		
geop...		
#co...		Legend color. The user selects the bounding b...
lege...		
#co...		Legend scale. The user selects the bounding b...
lege...		
#comment		These are parameter configurations probably ...
colorspace		
created		
validation		

Figure 10: *Annotation tool (AnT) user interface.*

5. From data to content

Environmental node discovery and data extraction services provide a large amount of relevant input data, which need to be assimilated in the context of a user-defined inquiry. Given that individual competing pieces of information from different sources can seldom be regarded as equally relevant, a procedure of fusion of competing or complementary data and their incorporation as content into the ontology is needed.

State-of-the-art proposals for data fusion from multiple resources have been successful when the data showed no major deviations (Hoek *et al.,* 2008; Janssen *et al.,* 2008; Weigel *et al.,* 2008). However, they fall short when the input data are extremely heterogeneous and contain simultaneous model forecasts and observations of varying reliability, time of validity and location – as it is the case in PESCaDO. Spatial and temporal gaps are also a matter of concern: there are only a finite number of measurement stations, and forecasting models have a finite spatial and temporal resolution. PESCaDO thus requires a fusion model that (a) facilitates the representation of a collection of environmental data extracted from different sources as one set of data, (b) assesses the quality of the data and chooses the best or merges independent data describing the same phenomenon such that the fusion outcome is of superior quality than that of the individual sources (Potempski & Galmarini, 2009) and (c) facilitates the extrapolation of small geographic and temporal gaps in the input data. This presupposes a general measure for information relevance and quality. All components of meteorological and AQ data reflect conditions of a certain place and time. These data components can be regarded as statistical estimations $\theta_p(r_p, t_p)$ for the conditions $\theta(r_0, t_0)$ in the user-defined area and time, where

$$\theta(r_0, t_0) = \theta_p\left(r_p, t_p\right) + \epsilon_p \tag{1}$$

(with ϵ_p as the error in terms of statistical variance and bias). The algorithm that is used for calculating the ensemble value requires information about the statistical properties of ϵ_p, namely, its variance $VAR[\epsilon_p]$. The Fusion Service (FS) estimates an aggregate statistical variance measure for each ϵ_p. These variance measures are then used for the assignment of averaging weights to each $\theta_p(r_p, t_p)$. A large estimated aggregate variance causes the assigned weight to decrease, while the data from the more accurate and relevant sources are assigned higher weights and, thus, gain more emphasis in the fusion.

In general, the procedure for fusing environmental data for a certain variable with respect to user-defined time and location consists of three stages: (a) evaluation of data variance; (b) optimal data weight calculation and merge; and (c) bias correction. Because in PESCaDO the fusion needs to be performed for a large set of variables over a specified period, this procedure is iterated several times to produce complete hourly fused time series for each variable while keeping the location fixed. For instance, if FS retrieves information from five different variable types describing conditions in a 24-h interval, then 5×24 fusion iterations would be performed. The fused data are fed into the KB of the system.

5.1. Data variance estimation

The variance of ϵ_p, or $VAR[\epsilon_p]$, is affected by the capability of the information source to properly assess the phenomenon of interest, for example, the sensor/model quality and/or its operational capacity. In addition, it is well known that information about air pollutant concentrations and weather conditions loses accuracy rapidly as a function of its ageing because atmospheric conditions are ephemeral and volatile phenomena, affected by randomness and entropy. Furthermore, a data point near the location r_0 should always gain more emphasis in the fusion in contrast to other data points that describe the conditions at more remote locations. Thus, we assume that the variance related to ϵ_p is the sum of these three individual (independent and thus summable) components, given by

$$VAR\left[\epsilon_p\right] \approx f(d) + g(\tau) + VAR\left[\theta_p\right] \tag{2}$$

where $f(d)$ is the variance component as a function of d, $g(\tau)$ is the temporal variance component as a function of τ, and d and τ are defined as follows:

$$d = \| r_0 - r_p \|, \quad \tau = \left| t_0 - t_p \right| \tag{3}$$

$VAR[\epsilon_p]$ in equation (2) describes the native quality of the information source in terms of variance or, in other words, the capability to estimate $\theta(r_0, t_0)$ when d and τ are equal to zero. For the evaluation of $VAR[\theta_p]$, stored information about the source's prediction accuracy in the past can be used, evaluated by the Uncertainty Metrics Tool depicted in Figure 11.

The PESCaDO framework allows for an imprecise definition of the location r_p for the estimator θ_p. This is usually the case, for instance, with extracted weather forecasts for cities. In these cases, r_p actually pinpoints the center of the city while information represents the conditions throughout the city. Then, the coordinates are flagged as approximations and set $d_p = min\{r_c, \| r_0 - r_p \|\}$, where r_c is the radius of the city.

The variance models $f(d)$ and $g(\tau)$ can be formulated in terms of statistical methods. In the FS, the methods have been formulated individually for each air pollutant using regression analysis with historical measurement data. For the pilot application in PESCaDO, these data represent 6 to 43 measurement stations across Finland, depending on the measured phenomenon. More specifically, the following simple regression models have been employed:

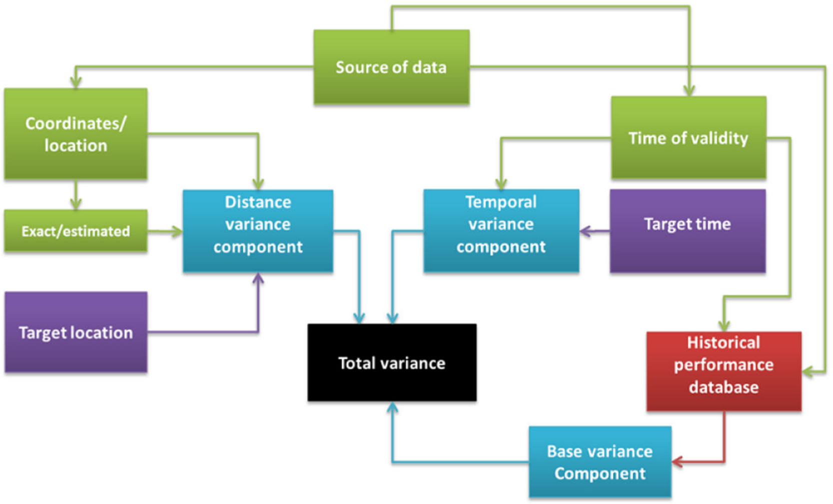


Figure 11: *Impresion estimation tool in* PESCaDO.

$$g(\tau) = a_3\tau^3 + a_2\tau^2 + a_1\tau + a), \quad f(d) = b_1 d \qquad (4)$$

where the parameters $a_3 \ldots a_1, a$ and b_1 are defined by statistical regression techniques. The achieved correlation of $g(\tau)$ polynomial models is generally very high for the domain of interest ($\tau < 36h$). In the formulation of $f(d)$, the capability of a measurement station to predict the measured phenomenon at a distance of d (covariance of the two time series) is evaluated. For illustration, see Figure 12 for the calculated average variances of measurement station pairs as a function of distance in the case of fine particles (PM2.5) for Finland. The derived parameterization of the regression curve, without the constant, can be used to represent $f(d)$.

5.2. *Optimal weight calculation*

Assuming all data sources to be independent and non-biased, an optimal ensemble value $\theta_F(r_0, t_0)$ can be calculated according to Potempski and Galmarini (2009) as

$$\theta_F(r_0, t_0) = \Sigma_{p=1}^{n} w_p \theta_p(r_p, t_p) \qquad (5)$$

where the individual weight w_p is given by

$$w_p = VAR[\epsilon_p]^{-1} / \Big(\Sigma_{p=1}^{n} VAR[\epsilon_p]^{-1} \qquad (6)$$

To assure statistical independence of $\theta_1 \ldots \theta_n$, only the most relevant estimator θ_p per data source is selected for the ensemble value calculation in equation (5). If a collection of estimators $\{\theta_1(r_1, t_1), \ldots \theta_k(r_k, t_k)\}$ is available from the same source, the selected θ_p to represent the source is simply the one with the lowest $VAR[\epsilon_p]$ from the collection. In the particular case of extracted time series from measurement stations, the estimator which has the smallest τ is selected to represent the source, as d and the base variance are the same for all $\theta_1 \ldots \theta_k$. Theoretically, it can be shown that the fused ensemble value $\theta_F(r_0, t_0)$ is the optimal estimator in terms of the mean squared error and that the prediction accuracy increases while the number n of independent data sources is increased (Potempski & Galmarini, 2009). More importantly, $\theta_F(r_0, t_0)$ does not suffer from very low quality input data, as long as d and τ in equation (3) have been estimated reasonably well.

5.3. *Bias correction with geo-demographic profiling*

The user of PESCaDO may draw a polygon within a geographical area to request environmental information in this polygon. This implies the retrieval of all relevant data in the polygon. Usually, the polygon contains various different environments such as parks, suburbs and urban areas. The representativeness of the measured data for each of these environments varies significantly. Thus, while the measured data in a rural area can be extended to cover a wide area (of a up to 100 km) without major reservation, the data measured by an urban measurement station cannot (already 10 m away from the station, the concentrations can differ considerably). Therefore, the estimator $\theta_p(r_p, t_p)$ may very well be significantly biased with respect to $\theta(r_0, t_0)$ while having a low aggregate variance estimate. To be able to remove this bias due to the geographical irrelevance, the FS utilizes a geo-demographic profiling feature using CORINE land use data in conjunction with a population density data from 2010.[14]

Based on the surrounding land use, the expected hourly average concentration of a certain pollutant in any location can be evaluated fairly accurately with the help of six variables: (a) population count, (b) suburban land use, (c) urban land use, (d) industrial land use, (e) roads and (f) vegetation. The land use variables are equal in this case to

[14] CORINE (COoRdinate INformation on the Environment) is a pan-European land cover/land use map database for non-commercial use provided by the European Environmental Agency; http://www.eea.europa.eu/publications/COR0-landcover.

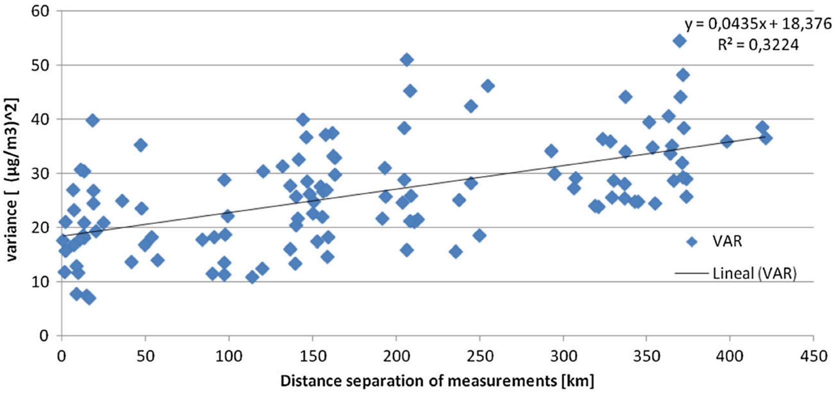


Figure 12: *The capability of a PM2.5 measurement station to predict another PM2.5 measurement as a function of distance (smaller is better). A data point describes the statistical variance between two PM2.5 measurement time series. The archived time series were measured in various locations in Finland in 2010.*

their relative frequency (percentage of total) inside the evaluation radius, which has been fixed to 200 m around the target location. To be able to distinguish larger cities from smaller ones, the evaluation radius for population count is much larger, approximately 6 km. In our experiments, in which a Monte Carlo simulation was used, both radia proved to result in the highest correlation of the expected average concentration (by the profiling feature) and the measured hourly average concentration. In Figure 13, the capability of the profiling tool to predict hourly average concentration of NO_2 in four different environments is demonstrated. The profiling tool can be calibrated or updated using the information that has flown through the PESCaDO system. Thus, FS can adapt and update the parameters of all statistical models.

By comparing the profiles of the target location and the specific location for input data (e.g., by using measurement station coordinates), it is possible to reduce the bias resulting from the differences between the expected hourly averages of air pollutant concentrations. Our latest evaluation studies suggest that when using the profiling feature set-on, the prediction error is at most test locations 30–60% less than it would be without the profiling feature.

6. Processing the content for the user

Once the assessed and fused data have been incorporated, the processing of the content of the KB for the user consists of two major tasks: (a) interpretation of the environmental

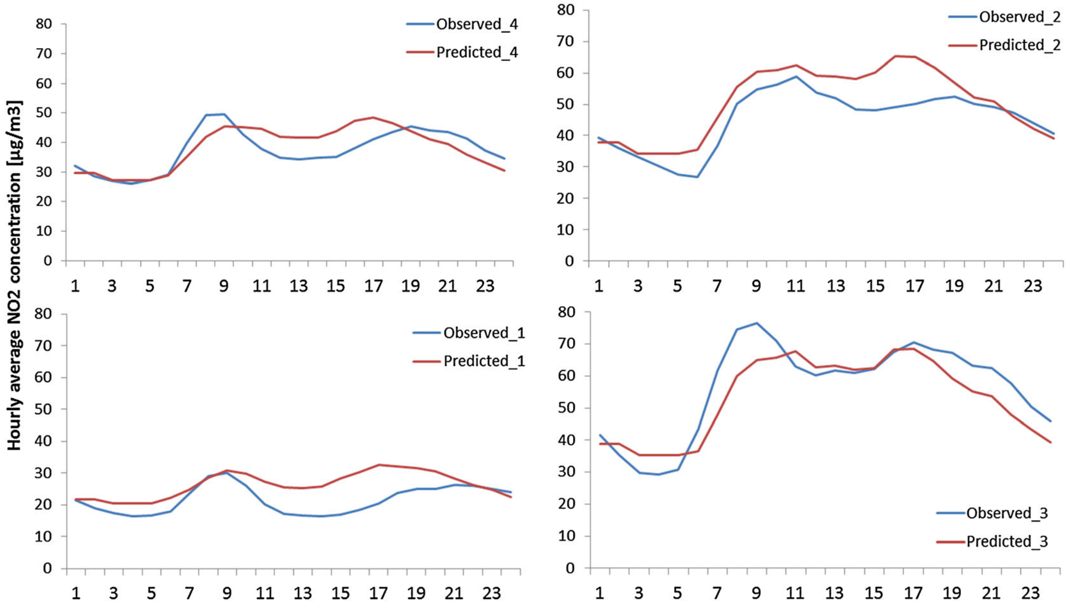
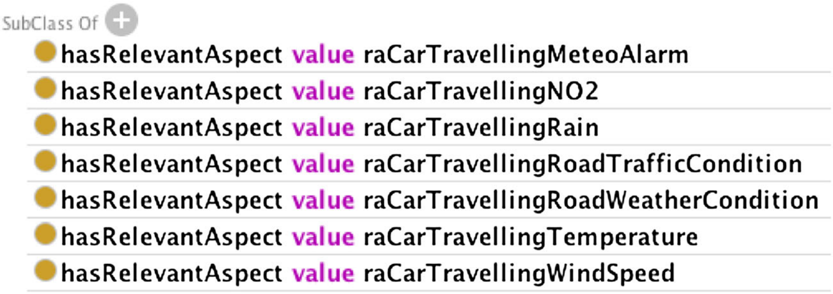


Figure 13: *Observed and estimated hourly averages of NO_2 at four measurement sites (evaluation period January–May 2011). The predicted hourly NO_2 concentrations are based on multi-variable regression models, which have been derived from another set of measurement time series from 2010.*



SubClass Of

hasRelevantAspect value raCarTravellingMeteoAlarm
hasRelevantAspect value raCarTravellingNO2
hasRelevantAspect value raCarTravellingRain
hasRelevantAspect value raCarTravellingRoadTrafficCondition
hasRelevantAspect value raCarTravellingRoadWeatherCondition
hasRelevantAspect value raCarTravellingTemperature
hasRelevantAspect value raCarTravellingWindSpeed

Figure 14: *The 'hasValue' restrictions associated to the 'Car Travelling' activity, stating that information about road traffic, road weather, rain, NO_2, temperature, wind and meteo alarm is relevant to requests involving this activity.*

content in view of the profile and needs of the user and (b) production of personalized information.

6.1. Interpretation of the environmental content

In order to adequately interpret the environmental data in view of the profile and needs of the user, two subsequent tasks need to be addressed. Firstly, all and only the relevant data for the given user and given request have to be identified, and secondly, the selected data have to be processed in order to produce personalized content. These two tasks go significantly beyond the scope of the common use of ontologies for environmental data (see, e.g., Ceccaroni *et al.*, 2004).

For the first task, mappings have been encoded in the ontology between the three main modules, Request, User and Activity of the PESCaDO-PDL (Section 2), and the types of environmental data supported by the system, in order to associate PDL statements (and thus user profile and inquiry elements) with environmental data and subsequently be able to identify relevant data.

The mappings, which have been defined together with environmental and health experts, are formalized as OWL 'hasValue' restrictions on the classes of the problem partition of the ontology. For instance, a restriction of the form 'hasRelevantAspect hasValue Rain' on the class that characterizes the users sensitive to some pollen states that data about precipitation should be retrieved and taken into consideration when providing decision support for this type of user. Figure 14 shows the 'hasValue' restrictions associated with the class representing the activity of travelling by car in the PESCaDO ontology.

The definition of mappings such as 'hasValue' allows us to automatically determine the environmental data types for which data are to be retrieved via description logics (DL)-based reasoning, by checking the new assertions inferred by the OWL reasoner from the request, the user profile and activity individuals that form the processed user decision support request. Furthermore, it supports easy extension/adaptation to new problem types added to the PESCaDO-PDL or to new user profiles: only the relevant aspect mappings for the new request types/activity types/user profile classes have to be defined. The techniques for identifying the relevant data for the given user and given request are implemented in the *Relevant Aspects Computation Service*.

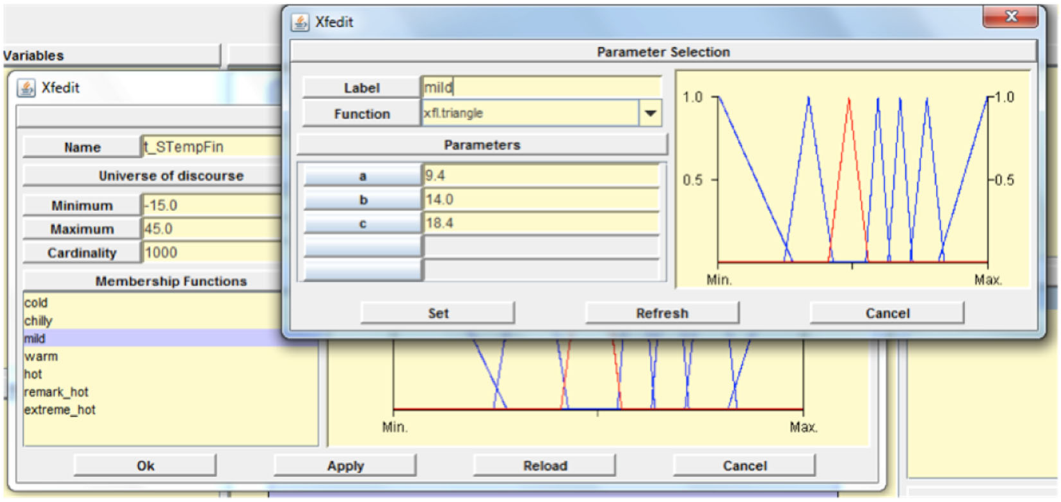
Once the raw environmental data that are relevant to the given user request have been selected, they are processed in order to produce personalized content. The personalized content comprises the following:

(1) Aggregations of raw data to favour the communication of environmental conditions to the users;
(2) Qualitative representation of a numerical value associated to a datum (e.g., the fact that a 5°C is considered a mild temperature);
(3) Detection of exceptional concentrations of air pollutants;
(4) Warnings or recommendations triggered by environmental conditions to support a user in making decisions (e.g., the fact that if the concentration of birch pollen is very high and the user is allergic to this pollen, a warning should be issued, and some suggestions on how to react should be proposed);
(5) Hints on possible causes of exceptional AQ episodes;
(6) LSRs connecting the facts described in the ontology (e.g., a relation stating that a certain data implies a warning).

For the computation of data aggregation, parameterized procedures have been defined that first query the PESCaDO KB to retrieve the data to be considered in the computation, then apply the appropriate aggregation function to the returned values and finally store in the ontology new environmental data representing the aggregated data.

For the computation of the qualitative representation of a numerical value, fuzzy reasoning has been applied. Fuzzy reasoning is based on the theory of fuzzy sets, which mimics human reasoning in its use of imprecise information to generate decisions (Zadeh, 1975).[15] An element of a fuzzy set belongs to the set in a degree of membership, a real number from the [0,1] interval. Examples are 'hot temperature', 'low pressure', 'satisfactory AQ index', and so on. A fuzzy set is formally defined by a membership function, which maps the universe of the discourse to the

[15] In PESCaDO, fuzzy reasoning is performed by exploiting Xfuzzy (http://www2.imse-cnm.csic.es/Xfuzzy/), a development environment for fuzzy logic systems.



Xfedit

Parameter Selection

Label: mild

Function: xfl.triangle

Parameters

a: 9.4

b: 14.0

c: 18.4

Variables

Xfedit

Name: t_STempFin

Universe of discourse

Minimum: -15.0

Maximum: 45.0

Cardinality: 1000

Membership Functions

- cold
- chilly
- mild
- warm
- hot
- remark_hot
- extreme_hot

Figure 15: *Membership function for 'mild summer temperature' in the fuzzy set for Finland.*

[0,1] interval. An example of a membership function for the fuzzy set 'mild summer temperature in Finland' is given in Figure 15. The interpretation of this function is that when the summer temperature is (more or less) between 10 and 18, we call it *mild*; close to the left border of the above interval, it becomes in between *chilly* and *mild*, while close to the right border, it is in between *mild* and *warm*; around the middle of the interval the 'mildness' of the temperature is maximal with the value of membership function close or equal to 1. The function type in the given example is triangular, which is one of the most commonly used membership function types; other widely used examples are Gaussian ('bell') and trapezoid functions.

For the detection of air pollutant concentrations above legally defined thresholds, a highly modular ontology-based computation has been implemented. All exeedances (together with their relevant details such, e.g., the threshold value) are formally represented in the PESCaDO ontology as instances of the class 'ExceedanceType' (or one of its subclasses). The system queries (using SPARQL) the PESCaDO ontology for the supported exceedance types, retrieving all individuals that are instances of the class 'ExceedanceType'. All parameters of each exceedance type are retrieved by the assertions made on the corresponding ontology instance. The system then calls the appropriate parameterized procedure (to which all the parameters of the exceedance type individual are passed as input values) to detect actual exceedances among the data stored in the PESCaDO KB.

For the generation of warnings and recommendations, the suggestion of possible causes of exceptional AQ episodes, and the instantiation of LSRs between the facts in the PESCaDO KB corresponding to the current request, we combined DL reasoning with rule-based reasoning. For this purpose, a two-layer reasoning infrastructure has been realized. The first layer exploits the HermiT reasoner (Shearer *et al.*, 2008) for the OWL DL reasoning services. The second layer is stacked on the top of the previous layer and implements the Jena RETE rule engine,[16] which performs the rule-based reasoning computation. Figure 16 shows an example of a rule for triggering the introduction of a recommendation to pollen-sensitive users in case of abundant pollen levels.

The produced personalized content is put back into the PESCaDO KB, stored as a set of instances and assertions on them. This enables us to relate the personalized content generated by the system with the data and the request description that were processed to produce it. The techniques for producing personalized content for the given user and given request are implemented in the Decision Service (DS).

6.2. *Personalized information production*

The generation of personalized information from the content provided by the DS is accomplished in PESCaDO using two modi: the text mode and the graphical mode, or a combination of both. The choice of the modi for the communication of a specific information is handled by a rule-based module and depends on the following:

- The profile of the user and the nature of the query; for example, if the user is a citizen asking for decision support in planning of the weekend in a national park, the text mode is chosen to be the main mode and the graphical mode as the complementary mode.
- The nature of the information to be communicated; for example, meteorological information suggests the use of graphical elements, while high concentrations of air pollution accompanied by health warnings and suggestions suggest the text mode.

[16] Jena – A Semantic Web Framework for Java. http://jena.sourceforge.net/index.html.

```
[ruleAbundantPollen:
(?request rdf:type pescadoProblem:AnyHealthIssue)
  (?request pescadoProblem:hasUser ?user)
  (?user pescadoProblem:isSensitiveTo ?pollen)
  (?pollen rdf:type pescadoData:PollenDataType)
  (?request pescado:hasGeoArea ?geoArea)
(?request pescado:hasData ?data)
(?data pescadoData:hasEnvironmentalDataType ?pollen)
(?data pescadoData:hasAggregationType pescadoData:max)
(?data pescadoData:hasRating ?rating)
(?rating pescadoData:hasRatingValue pescadoData:abundantPollen)
makeTemp(?rec)
   ->
(?rec rdf:type pescadoConclusions:Recommendation)
(?rec pescadoConclusions:hasRecommendationType
          pescadoConclusions:recommendation_abundantPollen)
(?rec pescadoConclusions:hasWeight 1.0xsd:double)
(?request pescado:hasConclusion ?rec)
(?data pescado:ProduceConclusion ?rec)]
```

Figure 16: *Example of a rule for triggering the introduction of a recommendation to pollen-sensitive users in case of abundant pollen levels.*

- The type of the requested service: Web-oriented pull service (which favours the graphical mode) or, for example, Twitter-based push service (which requests the text mode).

6.2.1. Multilingual textual environmental information generation PESCaDO's textual information generation module is a fully fledged multilingual language text generator in that it covers all three major tasks of text generation: (a) selection of the content that is to be communicated to the user; (b) discourse structuring (DiS) of the selected content; and (c) linguistic generation. In automatic report generation in general and in environmental information generation in particular, most often template-based generation has been used, drawing upon predefined content and discourse schemata and more or less rigid sentence templates, which are filled with actual data in the course of generation; see, for example, Busemann and Horacek (1997), Coch (1998) and Bohnet *et al.* (2001) for environmental information generators and Sripada *et al.* (2003), Yu *et al.* (2007) and Portet *et al.* (2009) for generators in other fields. In PESCaDO, the personalization of information and the great variety of linguistic constructions require more flexible solutions.

6.2.1.1. Content selection

Content selection operates on the output of the DS. It selects the content to be included in the report and groups it by topic. As is common for robust state-of-the-art report generators, CS draws upon a number of schemas in the sense of McKeown (1985), where each schema models a specific theme and specifies the type of content elements that are to be communicated in the context of this theme. In PESCaDO, schemas on AQ, weather, pollen and so on are defined. Each schema is implemented as a SPARQL query. Thus, there is a SPARQL query for instantiating the schema on AQ and related information: minimum and maximum values and ratings of the AQ index, pollutant(s) that contribute(s) to the AQ index, conclusions (i.e. warnings and recommendations) for the AQ and any LSRs between the schema's components. Similarly, there is a SPARQL query for instantiating a schema for each type of pollen, with minimum and maximum counts and ratings, conclusions related to these counts, and any LSRs between the schema's components associated with the schema.

Figure 17 presents graphically a sample schema for birch pollen comprising the pollen rating, associated recommendation and the implication relation between the two.

The inclusion of a given element in a schema can be subject to some restrictions defined in the queries. For instance, if the minimum and maximum AQ index ratings are identical, or if the maximum AQ index rating has an associated conclusion, then only the maximum AQ index rating is selected (the minimum AQ index rating is omitted). Similarly, a recommendation is selected only if it has a weight of at least 80%, where a weight of 100% is for abundant pollen and a weight of 0% for no pollen, thus ensuring that only *important* conclusions are conveyed to the user.

6.2.1.2. Discourse structuring

In contrast to, for example, Sripada *et al.* (2003) and Portet *et al.* (2009), PESCaDO flexibilizes the rigidness of schemas, which, in their traditional interpretation, prescribe not only which content elements are to be communicated but also how they are grouped and in which order they are presented. This is achieved by the use of (a) *Elementary Discourse Units* (EDUs) into which related individual elements are assembled (to be interpreted later in the generation chain as propositional units) in accordance with specific criteria, (b) LSRs, which are projected onto discourse relations in the sense of the Rhetorical Structure Theory (Mann & Thompson, 1988) to form a tree-like discourse structure and (c) context-dependent ordering of EDUs.

Subsequently, PESCaDO's DiS is a pipeline of three rule-based submodules: (1) EDU determination; (2) LSR – discourse relation projection; and (3) EDU ordering. All three of them draw heavily on what is known as *domain communication knowledge* (Kittredge *et al.*, 1991): the expert knowledge how to structure the discourse in a given domain. The domain communication knowledge of the environmental domain has been acquired and formalized in terms of rules

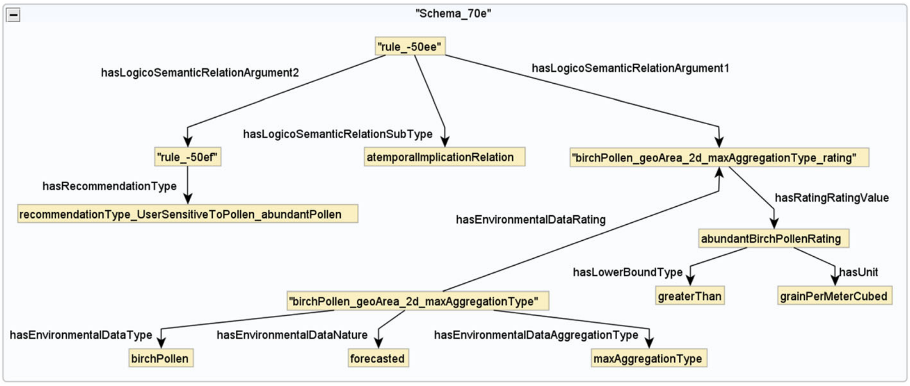

"Schema_70e"
"rule_-50ee"
hasLogicoSemanticRelationArgument2
hasLogicoSemanticRelationArgument1
hasLogicoSemanticRelationSubType
"rule_-50ef"
atemporalImplicationRelation
"birchPollen_geoArea_2d_maxAggregationType_rating"
hasRecommendationType
hasRatingRatingValue
recommendationType_UserSensitiveToPollen_abundantPollen
hasEnvironmentalDataRating
abundantBirchPollenRating
hasLowerBoundType
hasUnit
greaterThan
grainPerMeterCubed
"birchPollen_geoArea_2d_maxAggregationType"
hasEnvironmentalDataType
hasEnvironmentalDataNature
hasEnvironmentalDataAggregationType
birchPollen
forecasted
maxAggregationType

Figure 17: *Sample schema.*

and constraints discussed in the succeeding text in interviews with environmental experts. The output of the DiS module is a *text plan*, which serves as input to the linguistic generation module (cf. Figure 18 for illustration).

Let us discuss the aforementioned three submodules in turn.

(1) *EDU determination:* EDU determination groups topically related individuals into propositional units, starting from the schemas determined during CS. Figure 18 shows two EDUs identified in the pollen-related schema, one for the pollen rating and another for the message with the recommendation concerning the actions of the user in view of this rating. As already CS, EDU determination is handled by SPARQL queries. For instance, *AQ* query creates the EDU with the AQ index value; the *exceedance* query, the EDUs that capture the transgression of predefined thresholds by the measured/forecasted parameters; the *pollutant* query, the EDUs with the measurements of the individual air pollutants; the *missing data* query, the EDUs that communicate for which pollutants no measured/forecasted data are available for the requested time; and so on.

(2) *Mapping LSRs to discourse relations:* LSRs introduced between individuals of the ontology (Section 3.2) allow us the derivation of discourse relations and thus also the derivation of a tree-like discourse structure, which influences linguistic generation. Currently, we work with six LSR-to-discourse relation mapping rules, which proved to be sufficient for our application:

- `Non-volitional Cause` ⇒ `Circumstance`
  (as expressed, e.g., by *With the high levels of carbon monoxide, the air quality will be poor*).

- `Implication`, whose consequent is an exceedance ⇒ `Circumstance`
  (as expressed, e.g., by *With a value of 186 mg/m^3, the ozone threshold was exceeded*).

- `Implication`, whose consequent is a warning ⇒ `Elaboration`, whose satellite is the warning
  (as, e.g., in *The level of nitrogen dioxide was exceeded. Nitrogen dioxide causes respiratory symptoms, especially in children and asthmatics …*).

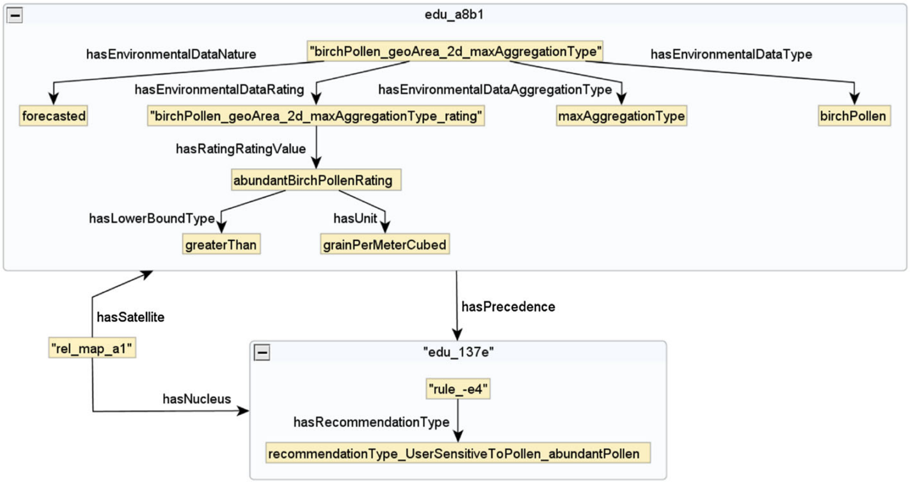

edu_a8b1
hasEnvironmentalDataNature
"birchPollen_geoArea_2d_maxAggregationType"
hasEnvironmentalDataType
hasEnvironmentalDataRating
hasEnvironmentalDataAggregationType
forecasted
"birchPollen_geoArea_2d_maxAggregationType_rating"
maxAggregationType
birchPollen
hasRatingRatingValue
abundantBirchPollenRating
hasLowerBoundType
hasUnit
greaterThan
grainPerMeterCubed
hasSatellite
hasPrecedence
"rel_map_a1"
"edu_137e"
"rule_-e4"
hasNucleus
hasRecommendationType
recommendationType_UserSensitiveToPollen_abundantPollen

Figure 18: *Sample text plan.*

- `Implication`, whose consequent is a recommendation ⇒ `Evidence`
  (as, e.g., in *Birch pollen is abundant. Most of the sensitive people have symptoms.*)

- `Implication` between a value and a rating ⇒ `Elaboration`
  (as, e.g., in *the maximum temperature will be low* ($-4^{\circ}C$)).

- `List` ⇒ `List`
  (as, e.g., in *With the high levels of ozone and carbon monoxide, the air quality will be poor*).

(3) *EDU ordering:* EDU ordering introduces precedence relations between obtained EDUs using a series of constraints that apply under specific conditions or that define the default order between EDUs. Consider, for example,

```
 IF    an EDU with the location and time of the
       user query is available
THEN   place this EDU first
or
IF     a measurement exeeds a legal threshold
       and the EDUs with the measurement and the
       threshold value are available
THEN   EDU-threshold < EDU-measurement as
```

examples of constraints of the first type and

```
EDU-temperature < EDU-wind speed < EDU-wind
direction < EDU-skies < EDU-humidity < EDU-
precipitation
```

as an example of constraints of the second type.

The definition of the problem of EDU ordering lends itself naturally to a resolution within the constraint satisfaction programming paradigm. We used the Java Choco API[17] and implemented the conditions for EDU ordering using simple SPARQL queries. The first solution offered by the solver is returned, provided it is consistent (i.e., each EDU has a position and each position is different); otherwise, the follow-up solutions are returned.

6.2.1.3. Linguistic generation

Our linguistic generation module is based on the extended multistratal linguistic model of the Meaning-Text Theory (Mel'čuk, 1988), such that generation consists of a series of transitions between structures of adjacent strata. The following strata are involved: (a) conceptual stratum, (b) semantic stratum, (c) deep-syntactic stratum, (d) surface-syntactic stratum, (e) deep-morphological stratum and (f) surface-morphological stratum. For each pair of adjacent strata S_i and S_{i+1}, a transition grammar G^i_{i+1} is defined such that any well-formed structure S_{i_j} of S_i is mapped by G^i_{i+1} onto a well-formed structure S_{i+1k} of S_{i+1}, with S_{i_j} and S_{i+1k} being equivalent with respect to their meaning.

Because the DiS module delivers as output an RDF-based text plan, and the linguistic generator requires as input a conceptual structure in the sense of Sowa (2000), the text plan is mapped onto the conceptual structure before linguistic generation is triggered.

The specifics of Meaning-Text Theory based linguistic generation are presented in Lareau and Wanner (2007) and further elaborated on in a series of publications, including (Wanner *et al.* (2010), Bouayad-Agha *et al.* (2012a) and Bouayad-Agha *et al.* (2012b). Therefore, let us simply illustrate in Figure 19 the kind of textual information produced by the linguistic generator in English, Finnish and Swedish (in this case, for a citizen upon the request of a comprehensive overview of environmental conditions). For contrast, consider the text in Figure 20 generated for an expert upon the request of information on AQ.

6.2.2. Graphical information generation

`PESCaDO` offers a set of Web-based visualization techniques that are capable of representing environmental data on a map. Some of the techniques are tailored to depict a specific data type, for example, particle flow for wind data, but most of them are multipurpose techniques. The visualization shows either continuous data, using overlay data views such as heatmaps, isolines, particle flow and graphs, or point data, using glyphs such as weather icons, bars and labels. The first always cover an area, while the second can be used, for instance, for a single location, for locations along a defined route, for a regular grid of locations in an area or for a separate view outside of the map.

Continuous data views depict the data from a whole area at a given point of time. If different time steps are relevant (e.g., throughout one day or multiple days), an animation can be used to give an overview, or the users can browse through the time steps manually using a slider control. A very common means are heatmaps. Heatmaps use colour to depict continuous data in a two-dimensional area. The heatmap is drawn semi-transparently over the map (Figure 21(a)), which allows for combining heatmaps with arbitrary maps and eliminating the need to redraw orientation guides such as state borders or city names. The conversion from data values to colour values is personalized to the current season and to each user, which is especially important for visually impaired users. Heatmaps are very suitable for showing a broad overview and are mainly applicable to dense data (in the case of sparse data, the interpolation between the actual data points may lead to false impressions). This can be countered by increasing the translucency of the heatmap if the uncertainty is high. Data intervals can also be shown in a heatmap using animation.

[17] http://www.emn.fr/z-info/choco-solver/

Situation in the selected area between 07/05/2012 (00h00) and 08/05/2012 (17h00).
The birch pollen count will be abundant. The air quality will be very poor, with the very high nitrogen dioxide and very high fine particles concentrations. The minimum temperature will be -1°C and the maximum temperature 9°C, the wind will be weak, the rain light and the sky condition between clear and partly cloudy.
Pollen recommendation: *Most of the sensitive people have symptoms.*
Air quality index recommendation: *Adverse health effects are possible on sensitive subpopulation.*
Nitrogen dioxide warning: *Nitrogen dioxide causes respiratory symptoms especially in children and asthmatics, because high concentrations of this gas cause contraction of the bronchial airways. It may increase the sensitivity of the airways to other irritants such as cold air and pollen.*
Fine particles recommendation: *Avoid excessive physical activity if you experience symptoms.*

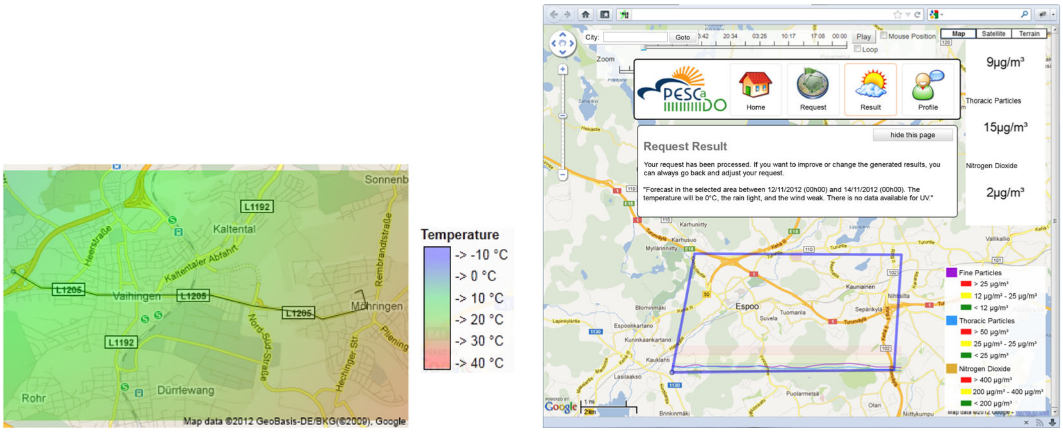
Olosuhteet valitulla alueella 7.5.2012 (klo 00:00) ja 8.5.2012 (klo 17:00) vlill.
Ilmassa on paljon koivun siitepölyä. Ilmanlaatu on erittäin huono erittäin korkeiden typpidioksidipitoisuuksien ja erittäin korkeiden pienhiukkaspitoisuuksien vuoksi. Alin lämpötila on -1°C ja ylin lämpötila 9°C, ja sää vaihtelee selkeästä puolipilviseen. Heikkoa tuulta. Heikkoa sadetta.
Siitepölysuositus: *Herkät ihmiset saavat oireita.*
Ilmanlaatuindeksisuositus: *Terveyshaitat ovat mahdollisia herkillä väestöryhmillä.*
Typpidioksidivaroitus: *Typpidioksidi lisää hengityselinoireita erityisesti lapsilla ja astmaatikoilla, koska se korkeina pitoisuuksina supistaa keuhkoputkia. Typpidioksidi voi lisätä hengitysteiden herkkyyttä muille ärsykkeille, kuten kylmälle ilmalle ja siitepölyille.*
Pienhiukkassuositus: *Vältä voimakasta rasitusta, mikäli saat oireita.*

Förhållandena på det valda området mellan 07/05/2012 (00h00) och 08/05/2012 (17h00).
Björkpollenhalten är hög. Luftkvaliteten är mycket dålig i och med den mycket höga kvävedioxidhalten och den mycket höga finpartikelhalten. Den lägsta temperaturen är -1°C och den högsta temperaturen 9°C och vinden är svag. Lätt regn. Klart eller svag molnighet.
Pollen: *De flesta känsliga personer kan ha allergiska symptom.*
Index för luftkvalitet: *Hälso-olägenheter är möjliga bland känsliga befolkningsgrupper.*
Kvävedioxid: *Kvävedioxiden ökar andningsorgansymptomer speciellt bland barn och astmatiker, eftersom den höga kvävedioxidhalten sammandrar luftrörer. Kvävedioxiden kan öka känsligheten för andra irritament, till exempel fr kall luft eller pollen.*
Finpartiklar: *Undvik hårda ansträngningar, ifall du får symptomer.*

Figure 19: *Multilingual environmental information produced in* `PESCaDO`.

The air quality will be fair, with the average thoracic particles concentration and the average concentration of nitrogen dioxide. There are no data available for sulfur dioxide.

Figure 20: *Environmental information produced in* `PESCaDO`.



Request Result

Your request has been processed. If you want to improve or change the generated results, you can always go back and adjust your request.

"Forecast in the selected area between 12/11/2012 (00h00) and 14/11/2012 (00h00). The temperature will be 0°C, the rain light, and the wind weak. There is no data available for UV."

(a) (b)

Figure 21: *(a) Heatmap data view; (b) graph and label data view.*

In the case of point data views, each glyph depicts exactly one data object; multiple glyphs can be combined to form glyph-based visualization. This has the advantage that each glyph in the view can depict a different point in time, eliminating the need for exploring a time range manually. For this purpose, an intuitive mapping from geo-coordinates to time is needed. A very common glyph type are labels (see the right-hand side of Figure 21(b)). The data value with its unit of measurement is placed at the specified geographical location. While placing the labels, it is important to avoid overlapping with other labels or lines from other visualizations. Therefore, the number of labels that can be used for the sampling of an area or route correlates with the label size.

All visualization types and available data types are registered in the Visualization Manager as map overlays and/or separate displays. Some of them supply additional control widgets for integration into the UI. Among the default control widgets are, for example, slider controls for the adjustment of the visualized point in time, the time range and the icon size. Individual visualizations can add their legends as custom control widget to the map interface. The Visualization Manager has his or her own control interface that allows the disabling/enabling of visualizations and

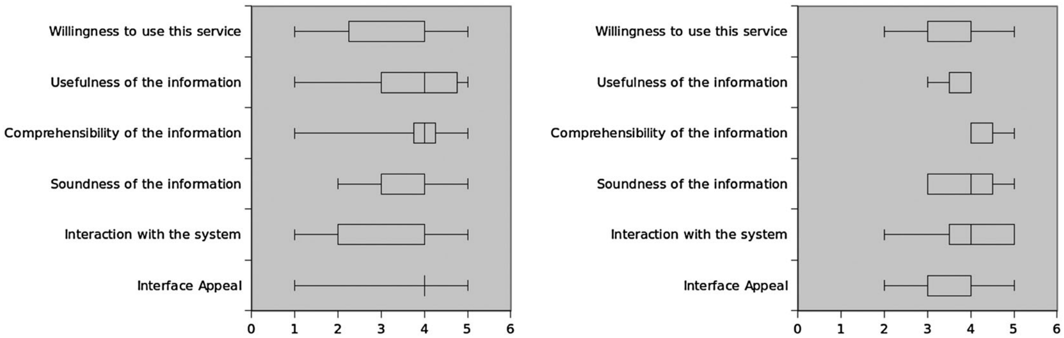


Figure 22: *Evaluation of* PESCaDO *by citizens (left) and environmental experts, when using* PESCaDO *for administrative decision support (right); in boxes that do not show the median marker, except in that for 'interface appeal' of experts, the median coincides with the right margin of the box; in the box for 'interface appeal' of experts, it coincides with the left margin of the box.*

controlling the mapping from data type to the visualization technique for further personalization.

As illustrated in Figure 21(b), the textual and graphical information are coherently integrated into the UI.

7. Evaluation of PESCaDO

The performance of the integrated PESCaDO system and of its individual modules has been thoroughly evaluated. Because the purpose of this paper is to present PESCaDO as an integrated system for the provision of personalized environmental information, we focus, in what follows, on the evaluation of the integrated system. The evaluation of the individual modules is presented elsewhere; see Tonelli *et al.,* (2011) and Moumtzidou *et al.* (2012a) for data acquisition, Johansson *et al.* (2013) for data fusion, Rospocher *et al.* (2012) for ontology construction, Rospocher and Serafini (2012) for the interpretation of the environmental content for decision support, Bouayad-Agha *et al.* (2012a) for textual information production and Bosch *et al.* (2012) for the user interface.

The evaluation of the integrated PESCaDO system has been carried out by groups of citizens and experts. In total, 18 citizens and seven environmental information experts (employees of the Helsinki Region Environmental Services Authority (HSY) and the Finnish Meteorological Institute, who were not involved in the project before) participated in the evaluation. Both groups had access to the final demonstrator of the system and were asked to fill out questionnaires compiled by experts from HSY and Finnish Meteorological Institute.[18]

Both groups were asked to rate their consent to a number of statements on a scale from 1 (strongly disagree) to 5 (strongly agree) – including 'The interface looks appealing', 'The interaction with the system is straightforward', and 'The provided information is comprehensible'. The box plots in Figure 22 show the dispersion of the ratings. With respect to the appeal of the interface, there is *de facto* no dispersion among citizens, while among the experts, the lower quartile is at 3 (which means 'undecided about the appeal of the interface') and the upper quartile at 4 ('I agree that the interface is appealing'). Unsurprisingly, the experts find the interaction with PESCaDO easier than the citizens. What might be more surprising is that the soundness of the information is judged in the average higher by experts than by citizens. One might have expected that environmental experts will be more critical with an automatically generated information than naïve users. The average 'comprehensibility of information' grade by citizens is 3.88 and by experts 4.3, that is, in both cases rather high; the dispersion among the ratings is in both cases limited. As the upper quartile of the box plot for the usefulness of the information among citizens shows, citizens tended to find the information provided by PESCaDO even somewhat more useful than the experts. The willingness to use the PESCaDO service is among citizens slightly lower (3.6 in the average) than among experts (3.9 in the average), but in both cases, it is high for an experimental implementation.

For the evaluation of PESCaDO by environmental experts, two additional evaluation runs have been carried out for appropriateness and completeness – the first is to assess the performance of PESCaDO when determining the relevant environmental data for a given request submitted to the system and the second is to assess the performance of the system when producing targeted personalized content (e.g., whether it is appropriate to say for −26.3°C that it is 'extremely cold' and whether the health warnings are included in the case of a high concentration of ozone).[19] In both runs, the evaluation was based on three different user scenarios that the experts had to have in mind: (a) a citizen with no detailed environmental background plans an excursion; (b) environmental administration staff member inquires about the measurements of air pollutants at a certain time point, and, in particular, whether any

[18] Interested citizens can inspect the Final PESCaDO Demonstrator at http://pescado-project.upf.edu/ and send their comments to the Consortium.

[19] The linguistic and visual quality of the produced information has been evaluated in the scope of the evaluation of the textual and graphical information generation modules.

pollutants are above or near the legal thresholds; and (c) an administrative staff member desires to set up an environmental service.

In the first run, the experts were presented in sequence the description of each of the scenarios and the list of environmental data determined as relevant for the scenario in question by the system and were asked to judge the appropriateness of these data in the context of the scenario. The outcome of the evaluation revealed an average appropriateness of 94% (with a standard deviation of the agreement between experts of 11%) and an average completeness of 92% (with a standard deviation of 8%). The appropriateness and completeness figures of all experts and all scenarios are most often distributed between 92% and 100% for both appropriateness and completeness. The median value is 100% for appropriateness and 92% for completeness.

In the second run, the experts were presented for each scenario the textual description of the scenario, a spreadsheet containing the relevant environmental data retrieved by the DS and the environmental information delivered by the system. The obtained results show an average appropriateness of 90% (with a standard deviation of 25%) and an average completeness of 87% (with a standard deviation of 23%) of the personalized content produced by the system. With the exception of a few outliers, the appropriateness and values obtained for all experts, all scenarios and all questions are, as a rule, distributed between 95% and 100%, while the completeness figures are mostly distributed between 75% and 100%. The median value is 100% in both cases.

8. Conclusions

The increasing need for advanced solutions in the area of personalized environmental information is generally acknowledged by experts and confirmed by polls among citizens. Thus, while in a Gallup poll from 2010 by HSY, 47% of the Finnish interviewees of age between 35 and 49 years, 57% between 25 and 34 years and 70% between 15 and 24 years considered a service such as PESCaDO useful and were ready to use it; in a poll from 2012, already 76% between 35 and 49 years old, 68% between 25 and 34 years old, and 75% between 15 and 24 years expressed this opinion. Providers of Web-based information are aware of this need and offer meteorological, AQ and pollen information via a multitude of web pages, which are intensely accessed. Unfortunately, the quality of this information varies significantly, such that it is difficult to rely upon one specific source without having carried out a thorough empirical evaluation study. Despite this, all state-of-the-art proposals use data from one source (e.g., one measurement network) and provide the same information for all users. In this paper, we argued that this is not appropriate. An appropriate solution ought, on the one hand, to take advantage of the wealth of the environmental data available in the Web and, on the other hand, tune the information to the needs of the respective user. The PESCaDO service has been developed with these two goals in mind. It offers integrated cutting edge Artificial Intelligence technologies for a series of tasks, including environmental data acquisition, processing and delivery to offer multilingual personalized information to various types of users. This makes it novel compared with the state-of-the-art user-oriented environmental services such as Busemann and Horacek (1997), Coch (1998), Bohnet *et al.* (2001), Bøhler *et al.* (2002) and Wanner *et al.* (2010), which tend to focus only on a subset of the tasks, for example, report generation as Busemann and Horacek (1997), Coch (1998), Bohnet *et al.* (2001) and Wanner *et al.* (2010) or raw data delivery as Bøhler *et al.* (2002). Those approaches that focus on report generation operate on raw data, which severely limits the potential of their rule-based inference engines applied to deduce the information that is to be communicated to the user. In contrast, PESCaDO operates on ontological representations and uses fully fledged reasoning engines. Furthermore, while PESCaDO uses an equally fully fledged report generator, Busemann and Horacek (1997), Coch (1998) and Bohnet *et al.* (2001) are based on template-based generation, which lacks the flexibility required by a personalized information production service.

However, it should be also clear that a number of challenges await to be solved to make PESCaDO a mature service – including, for instance,

(i) Environmental data extraction from web pages in any format in order to ensure that it is applicable to any geographical region and any unseen environmental node;
(ii) Broad coverage ontology learning that would allow us to acquire additional background knowledge on environmental, medical and social topics that might be relevant to the provision of personalized environmental information – without cost-intensive manual labour;[20]
(iii) Decision support models that are universal enough to cope with unexpected user inquiries;
(iv) Broad coverage robust multilingual generators that would take any well-formed conceptual structure as input and produce a report in a great variety of languages; and, finally,
(v) User privacy and personal data security: as a prototypical service, PESCaDO does not take any measures to ensure that sensitive user data such as domicile, life style preferences and allergies are handled in compliance with legal data protection regulations.

[20] Incorporation of, e.g., TaToo techniques developed specifically for the acquisition of environmental content (Schimak *et al.*, 2010) would help ease the knowledge bottleneck.

References

Bøhler, T., K. Karatzas, G. Peinel, T. Rose and R. San José (2002) Providing multi-modal access to environmental data-custamisable information services for disseminating urban air quality information in APNEE, *Computers, Environment and Urban Systems*, **26**, 39–61.

Bohnet, B., L. Wanner, R. Ebel, B. Knörzer, M. Tauber and W. Weiß (2001) Autotext-UIS: Automatische Produktion von Ozonkurzberichten im Umweltinformationssystem Baden-Württemberg, in *Proceedings of the Workshop Hypermedia und Umweltschutz*.

Bosch, H., D. Thom, G.-A. Heinze, S. Wokusch and T. Ertl (2012) Dynamic ontology supported user interface for personalized decision support, in *Proceedings of the 5th International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2012)*, IARIA, 101–107.

Bouayad-Agha, N., G. Casamayor, S. Mille, M. Rospocher, H. Saggion, L. Serafini and L. Wanner (2012a) From Ontology to NL: Generation of multilingual user-oriented environmental reports, in *Natural Language Processing and Information Systems: Proceedings of 17th International conference on Applications of Natural Language Processing to Information Systems (NLDB 2012)*, volume 7337 of *Lecture Notes in Computer Science*, 216–221, Heidelberg: Springer Verlag.

Bouayad-Agha, N., G. Casamayor, S. Mille and L. Wanner (2012b) Perspective-oriented generation of football match summaries: old tasks, new challenges. *ACM Transactions on Speech and Language Processing (TSLP)*, **9**(2), 3–31.

Busemann S. and H. Horacek (1997) Generating air-quality reports from environmental data, in *Proceedings of the DFKI Workshop on Natural Language Generation*, 15–21.

Ceccaroni, L., U. Cortés and M. Sànchez-Marrè (2004) OntoWEDSS: augmenting environmental decision-support systems with ontologies. *Environmental Modelling & Software*, **19**, 785–797.

Chang, C-C. and C.-J. Lin (2011) LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1–27. Software Available at: http://www.csie. ntu.edu.tw/ cjlin/libsvm.

Coch, J. (1998) Interactive generation and knowledge administration in MultiMeteo, in *Proceedings of the Ninth International Workshop on Natural Language Generation*, 300–303.

Epitropou, V., K. Karatzas, A. Bassoukos, J. Kukkonen, and T. Balk (2011) A new environmental image processing method for chemical weather forecasts in Europe, in *Proceedings of the 5th International Symposium on Information Technologies in Environmental Engineering*, P. Golinska, M. Fertsch and J. Marx-Gomez (eds), Berlin/New York: Environmental Science and Engineering, Springer Series, 781–791.

Hoek, G., R. Beelen, K. Hoogh, D. Viennau, J. Gulliver, P. Fischer and D. Birggs (2008) A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, **42**:7561–7578.

Janssen, S., D. Gerwin, F. Fierens, and C. Mensink (2008) Spatial interpolation of air pollution measurements using CORINE land cover data. *Atmospheric Environment*, **42**, 4884–4903.

Johansen, P.H., K. Karatzas, J.E. Lindberg, G. Peinel, and T. Rose (2001) Citizen-centered information dissimination with multimodal information channels. Sustainability in the information society, in *Proceedings of the 15th International Symposium on Informatics for Environmental Protection*.

Johansson, L., A. Karppinen and L. Wanner (2013) The fusion of meteorological- and air quality information for orchestrated services using environmental profiling, in *Proceedings of the Fusion 2013 Conference*, Istanbul, Turkey.

Karatzas, K. (2007) State-of-the-art in the dissemination of AQ information to the general public, in *Proceedings of the EnviroInfo Conference, Volume 2*.

Kittredge, R., T. Korelsky and O. Rambow (1991) On the need for domain communication knowledge. *Computational Intelligence*, **7**, 305–314.

Klein, T., J. Kukkonen, A. Dahl, E. Bossioli, A. Baklanov, A. Fahre Vik, P. Agnew, K. Karatzas, and M. Sofiev (2012) Interactions of physical, chemical and biological weather calling for an integrated assessment, forecasting and communication of air quality. *AMBIO*, **48**, 851–864.

Kumar, V. (2008) Soft computing approaches to uncertainty propagation in environmental risk management. Ph.D. thesis, Universitat Rovira i Virgili, Tarragona, Spain.

Lareau, F. and L. Wanner (2007) Towards a generic multilingual dependency grammar for text generation, in *Proceedings of the Grammar Engineering Across Frameworks Workshop (GEAF07)*, T. King and E.M. Bender (eds), Stanford, CA: CSLI, 203–223.

Li, J., G.H. Huang, G. Zeng, I. Maqsood and Y. Huang (2007) An integrated fuzzy-stochastic modelling approach for risk assessment of groundwater contamination, *Journal of Environmental Management*, **82**, 173–188.

Mann, W.C. and S.A. Thompson (1988) Rhetorical structure theory: toward a functional theory of text organization, *Text*, **8**, 243–281.

McKeown, K.R. (1985) *Text Generation*, Cambridge: Cambridge University Press.

Mel'čuk, I.A. (1988) *Dependency Syntax: Theory and Practice*, Albany: SUNY Press.

Moßgraber, J. and M. Rospocher (2012) Ontology management in a service-oriented architecture, in *Proceedings of the 23rd International Workshop on Database and Expert Systems Applications – DEXA 2012*. ISBN: 978-0-7695-4801-2.

Moumtzidou, A., V. Epitropou, S. Vrochidis, S. Voth, A. Bassoukos, K. Karatzas, J. Mossgraber, I. Kompatsiaris, A. Karppinen and J. Kukkonen (2012a) Environmental data extraction from multimedia resources, in *Proceedings of the 1st ACM international workshop on Multimedia analysis for ecological data (MAED 2012)*, 13–18.

Moumtzidou, A., S. Vrochidis, S. Tonelli, I. Kompatsiaris and E. Pianta (2012b) Discovery of environmental nodes in the Web, in *Proceedings of the 5th IRF Conference*.

Munzner, T. (2012) Visulization principles, in *Cytoscape Symposium on Network Visualization,* http://www.cs.ubc.ca/tmm/talks/vizbi11/networkbio12.pdf.

Oyama, S., T. Kokubo and T. Ishida (2004) Domain-specific web search with keyword spices. *IEEE Transactions on Knowledge and Data Engineering*, **16**, 17–27.

Park, K. (2011) Modeling uncertainty in metric space, Ph.D. thesis, Department of Energy Resources Engineering, Stanford, CA: Stanford University.

Peinel, G., T. Rose and R. San José (2000) Customized information services for environmental awareness in urban areas, in *Proceedings of the Seventh World Congress on Intelligent Transport Systems*.

Pianta, E. and S. Tonelli (2010) KX: a flexible system for key phrase extraction, in *Proceedings of SemEval 2010*.

Portet, F., E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer and C. Sykes (2009) Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, **173**, 789–916.

Potempski, S. and S. Galmarini (2009) Est modus in rebus: analytical properties of multi-model ensembles, *Atmospheric Chemistry and Physics*, **9**, 9471–9489. doi:10.5194/acp-9-9471-2009.

Preece, J., Y. Rogers and H. Sharp (2002) *Interaction Design: Beyond Human-Computer Interaction*, New York: Wiley.

Rospocher, M. and L. Serafini (2012) An ontological framework for decision support, in *Proceedings of the 2nd Joint International Semantic Technology Conference (JIST2012)*.

ROSPOCHER, M., S. TONELLI, L. SERAFINI and E. PIANTA (2012) Corpus-based terminological evaluation of ontologies, *Applied Ontology*, **7**, 429–448.

SCHIMAK, G., A.E., RIZZOLI, G. AVELLINO, T. PARIENTE LOBO, J.M. F. LOPEZ and I. ATHANASIADIS (2010) Information enrichment using TaToo's semantic framework, *Communications in Computer and Information Science*, **108**, 149–159.

SHEARER, R., B. MOTIK and I. HORROCKS (2008) HermiT: a highly-efficient OWL reasoner, in *Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008 EU)*.

SOWA, J. (2000) *Knowledge Representation*, Pacific Grove, CA: Brooks Cole.

SRIPADA, S., E. REITER and I. DAVY (2003) SumTime-Mousam: configurable marine weather forecast generator. *Expert Update*, **6**, 4–10.

TANG, T.T., D. HAWKING, N. CRASWELL, and R. S. SANKARANARAYANA (2004) Focused crawling in depression portal search: a feasibility study, in *Proceedings of the 9th Australasian Document Computing Symposium,* http://www.people.eng.unimelb.edu.au/ammoffat/adcs2004/papers/paper01.pdf.

TONELLI, S., M. ROSPOCHER, E. PIANTA and L. SERAFINI (2011) Boosting collaborative ontology building with key-concept extraction, in *Proceedings of the IEEE Fifth International Conference on Semantic Computing (ICSC-2011)*.

USLÄNDER, T. (2007) Reference model for the ORCHESTRA Architecture Version 2.1, Technical Report OGC Best Practices Document 07–097, http://www.portal.opengeospatial.org/files/?artifact_id=23286.

USLÄNDER, T. (2009) Specification of the Sensor Service Architecture, Version 3.0 (Rev. 3.1), Deliverable D2.3.4 of the European Project SANY, FP6-IST-033564, Technical Report OGC Discussion Paper 09-132r1, http://www.portal.opengeospatial.org/files/?artifact_id=35888&version=1.

VROCHIDIS, S., H. BOSCH, A. MOUMTZIDOU, F. HEIMERL, T. ERTL and I. KOMPATSIARIS (2012a) An environmental search engine based on interactive visual classification, in *Proceedings of the 1st ACM international workshop on Multimedia analysis for ecological data (MAED 2012)*, 49–52.

VROCHIDIS, S., V. EPITROPOU, A. BASSOUKOS, S. VOTH, K. KARATZAS, A. MOUMTZIDOU, J. MOSSGRABER, I. KOMPATSIARIS, A. KARPPINEN and J. KUKKONEN (2012b) Extraction of environmental data from on-line environmental information sources, in *Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology, Volume 382. 3rd Intelligent Systems for Quality of Life information Services Workshop (ISQL 2012)*, 361–370.

WANNER, L., B. BOHNET, N. BOUAYAD-AGHA, F. LAREAU and D. NICKLAß (2010) MARQUIS: generation of user-tailored multilingual air quality bulletins, *Applied Artificial Intelligence*, **24**, 914–952.

WEIGEL, A.P., M.A LINIGER and C. APPENZELLER (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, **134**, 241–260.

WITTEN, I.H., E. FRANK, and M.A. HALL (2011) *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn, San Francisco: Morgan Kaufmann.

YU, J., E. REITER, J. HUNTER and C. MELLISH (2007) Choosing the content of textual summaries of large time-series data sets, *Natural Language Engineering*, **13**, 25–49.

ZADEH, L. (1975) Fuzzy logic and approximate reasoning, *Synthese*, **30**, 407–428.

ZAGORULKO, Y. and G. ZAGORULKO (2010) Ontology-based approach to development of the decision support system for oil-and-gas production enterprise, in *Proceedings of the 2010 Conference on New Trends in Software Methodologies, Tools and Techniques.*

The authors

Leo Wanner

Prof. Leo Wanner holds a diploma in Computer Science from the University of Karlsruhe and a PhD in Computational Linguistics from the University of The Saarland. Currently, he is an ICREA research professor at the Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona. Leo's research interests cover a broad range of areas in Computational Linguistics, including multilingual text generation, summarization, parsing, information extraction, computational lexicography, and computer assisted language learning. He has an extensive publication record and served on the program and organization committees of numerous conferences in the field.

Harald Bosch

Harald Bosch is with the Institute for Visualization and Interactive Systems at the University of Stuttgart, where he is currently pursuing a doctoral degree. He received a Master of Science in Information Systems from the University of Stuttgart and the Universität Hohenheim. His research interests comprise the fields of information visualization, human computer interaction, and visual analytics with a special focus on text-documents and semantic web.

Nadjet Bouayad-Agha

Dr. Nadjet Bouayad-Agha has been a researcher in the TALN group at the Universitat Pompeu Fabra (Barcelona, Spain) since 2002. She received her PhD in 2001 from the University of Brighton (UK) where she had been working as a research assistant then fellow since 1997. Her research interests include Discourse planning for Natural Language Generation from time-series and/or Semantic Web data, text simplification and paraphrasing in general, and specialized domains, leveraging corpora for knowledge acquisition for NLP tasks such as coreference resolution, or content selection.

Gerard Casamayor

Gerard Casamayor is a researcher in the Natural Language Processing Group at the Department of Information and Communication Technologies, Pompeu Fabra University. He holds a degree in Computer Science from the Polytechnic University of Catalonia (UPC) and a degree in Linguistics from Pompeu Fabra University. Gerard is a specialist in the use of natural language processing and Semantic Web technologies and has participated in a wide range of European research initiatives in these fields.

Thomas Ertl

Prof. Thomas Ertl received a Masters degree in computer science from the University of Colorado at Boulder and a PhD in theoretical astrophysics from the University of Tübingen. Currently, he is a full professor of computer science at the University of Stuttgart, Germany and Head of the Visualization and Interactive Systems Institute (VIS) and the Visualization Research Center of the University of Stuttgart (VISUS). His research interests include visualization, computer graphics, and human computer interaction in general with a focus on volume, flow, and particle visualization, parallel and hardware accelerated graphics, visual analytics of text collections and social media, user interfaces, and navigation systems for the blind. He is a coauthor of more than 400 scientific publications; he served on and chaired numerous committees and boards in the field.

Désirée Hilbring

Dr. Désirée Hilbring received her PhD in engineering (2005) from the University of Karlsruhe, Germany. She works in the Group ‘Architecture and Information Systems’ at the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB. Désirée gained experience with modern service oriented architectures and the semantic web in several European (ORCHESTRA, SANY, PESCaDO) and German national (EWS Transport, SiKomFan) research projects. Furthermore, she manages the development of the Bavarian version of a Water Information System, which supports the European Water Framework Directive (WRRL).

Lasse Johansson

Lasse Johansson, research scientist, MSc., is a PhD student, working on Atmospheric Dispersion Modelling at FMI since January 2011. He has a MSc. degree in Systems Analysis and Applied Mathematics from Aalto University in Finland. He is the lead developer and programmer for the FMI-ENFUSER service, which had a key role in the fusion of environmental information in the EU/PESCaDO (ICT-249584) project, as well as in CLEEN MMEA test-bed (TEKES/SHOK). He is also the supporting developer and lead programmer for the FMI-STEAM shipping emission model. Despite his relatively short academic career thus far, he has published several peer review publications and has given seven presentations at international scientific conferences. Furthermore, he has participated in several (six) internationally funded research projects.

Kostas Karatzas

Prof. Dr.-Eng. Kostas Karatzas is associate professor for Informatics Systems and Applications—Environmental Informatics at the Department of Mechanical Engineering, Aristotle University of Thessaloniki (AUTh), Greece, where he leads the Informatics Systems and Applications Group. Kostas holds a diploma and a doctoral degree in Mechanical Engineering and works mainly in the field of environmental informatics, urban environment management and information systems, environmental and energy data analysis, and forecasting with the aid of computational intelligence methods and mathematical models, participatory environmental sensing, and quality of life information services. He is an author and a co-author of more than 200 scientific publications and participated in several national and international research projects. Furthermore, he acts/has acted as consulting expert for the European Environment Agency, the Government of Cyprus, and the City of Thessaloniki.

Ari Karppinen

Dr.Tech, Doc, Ari Karppinen works as research manager at the Finnish Meteorological Institute. He is the author of approximately 300 scientific publications; 60 of these in refereed international journals. His research group (20 researchers) works on the development, evaluation, and application of atmospheric dispersion models.

loannis Kompatsiaris

Dr. Ioannis (Yiannis) Kompatsiaris is a senior researcher (researcher A’) with the Information Technologies Institute/ Centre for Research and Technology Hellas, Thessaloniki, Greece. His research interests include semantic multimedia analysis, indexing and retrieval, social media and big data analysis, knowledge structures, and reasoning and personalization for multimedia applications, eHealth and environmental applications. He received his PhD degree in 3D model-based image sequence coding from the Aristotle University of Thessaloniki in 2001. He is the co-author of 69 papers in refereed journals, 35 book chapters, eight patents and more than 240 papers in international conferences. He has been the co-organizer of various international conferences and workshops and has served as a regular reviewer for a number of journals and conferences. He is a senior member of IEEE and member of ACM.

Tarja Koskentalo

Lic. Tech. Tarja Koskentalo is the head of Air Protection Unit in Helsinki Region Environmental Services Authority HSY. She has a long experience in air quality issues and she is responsible for air quality monitoring, communications, and research in Helsinki metropolitan area. Tarja has co-authored more than 30 peer-reviewed scientific publications and participated in various research projects, for example, in the recent EU-funded projects SNOOP, PESCaDO and REDUST.

Simon Mille

Dr. Simon Mille is researcher in the Natural Language Processing Group at the Department of Information and Communication Technologies, Pompeu Fabra University, where he obtained his PhD in 2014. He is a specialist in symbolic multilingual natural language generation and in syntactic and semantic annotation of corpora. He has participated in various national and European projects involving data-to-text and text-to-text generation.

Jürgen Moßgraber

Dipl.-Inform. Jürgen Moßgraber holds a degree in Computer Science from the University of Karlsruhe, Germany. He is the head of the research group 'Architecture and Information Systems' at the Fraunhofer Institute of Optronics, System Technologies, and Image Exploitation IOSB. His research interests include the design and implementation of web-based information systems with state-of-the-art technologies. At the moment, his focus is mainly on utilizing research results from the areas of geographic information (conforming to OGC standards) and the semantic web (ontologies) and bringing them into real world systems. Jürgen has particular experience with the design of distributed systems handling large-scale databases for automated manufacturing systems and modern architectures for task-oriented surveillance systems with a service-based and event-based approach.

Anastasia Moumtzidou

Anastasia Moumtzidou received her diploma degree in Electrical and Computer Engineering in 2006, her first MSc degree dealing Advanced Systems of Computers and Communications in 2009, and her second MSc degree dealing with Informatics and Management in 2011, all from the Aristotle University of Thessaloniki. Since 2007, she has been working as a research associate in CERTH-ITI, and her research interests include software engineering for database systems and web-based applications, semantic multimedia analysis, and content-based image indexing and retrieval. She has participated in three European projects, and she is the co-author of more than 20 journal and conference publications.

Maria Myllynen

M. Sc. Maria Myllynen is an expert in air quality in Helsinki Region Environmental Services Authority HSY. She studied at the Faculty of Environmental Sciences, University of Eastern Finland. Later on, she also studied communication at the University of Helsinki and city planning at Aalto University. Maria's expertise is in air quality communications, and she has improved methods to inform the public on air quality. She also deals with air quality issues in city planning. She has been involved in EU-funded projects PESCaDO and REDUST.

Emanuele Pianta

Emanuele Pianta (in memoriam) graduated at the University of Padova in 1990 with a thesis on the Relevance Theory applied to the task of Automatic Language Generation. He worked as a research consultant at the University of Venice for 2 years on LFG-based Parsing and Generation, Computational Morphology, and teaching Logic Programming. In 1993, he joined ITC-irst (now FBK) as a tenure researcher. He worked on several national and international projects such as LRE-GIST, MultiWordNet, CStar-II, NESPOLE, MEANING, OntoText, PATExpert, Live Memories, LODE, PESCaDO, and TERENCE.

Marco Rospocher

Dr. Marco Rospocher is a research scientist at Fondazione Bruno Kessler (FBK). He received his PhD in Information and Communication Technologies from the University of Trento in 2006. His current research interests are in the area of Semantic Web and Knowledge Representation, focusing in particular on ontologies, formalisms for Knowledge Representation and Reasoning, and methodologies and tools for Knowledge Acquisition. He (co-)authored more than 50 papers in international journals, conferences, and workshops. He served as a program committee member in several international conferences, workshops, and PhD Symposiums, and he reviewed paper for several international journals. He has also been involved in a number of international research projects, including the EU-funded projects APOSDLE, PESCaDO, and NewsReader.

Luciano Serafini

Luciano Serafini is the head of the Data and Knowledge Management Research Unit at Fondazione Bruno Kessler, Trento, Italy. He has more than 25 years of experience in academic research and technology transfer in the area of knowledge representation and reasoning, artificial intelligence and Knowledge, and Data Management. His research interests include logic for distributed knowledge (since 1990 development of the logic of context) information integration, automated reasoning, multi agent system, ontological reasoning for the semantic web, collaborative knowledge aquisition and modelling, and integration of logical and statistical knowledge. Luciano has published more than 150 papers, and his h-index under Google Scholar is 37. He is a member of the PC committee of several top level conferences and workshops in these areas and contributes to many industrial, research, local, and European projects. He was the local chair of the 2002 edition of the European Summer School on logic language and information and the local co-chair of the 13th International Semantic Web Conference (2014) He regularly supervises Master and PhD students at the University of Trento, Verona, Milano and Graz, and his teaching activity includes courses in Database, Information systems, Mathematical Logic, Knowledge Representation and reasoning, and Semantic Web for the Master degree and Doctoral Degree at the University of Trento and Bolzano.

Virpi Tarvainen

Dr. Virpi Tarvainen is a senior scientist at the Atmospheric Dispersion Modelling group of the Finnish Meteorological Institute. She has obtained a PhD in Physics from the University of Helsinki. She has wide experience in modelling the physical and chemical processes in the atmosphere and biogenic volatile organic compound emissions of plants. She has also been a lecturer of atmospheric chemistry at the University of Helsinki for 10 years.

Sara Tonelli

Dr. Sara Tonelli is the head of the Digital Humanities research unit at Fondazione Bruno Kessler (FBK) in Trento. She got her Phd in Language Sciences at the University of Venice in 2010 and then she held a post-doc position in the Human Language technology group at FBK. Her main research interests include lexical semantics, frame semantics, and event-based text processing. She has been involved in several national and international projects such as PESCaDO, Terence, and NewsReader.

Stefanos Vrochidis

Dr. Stefanos Vrochidis received the diploma degree in Electrical Engineering from Aristotle University of Thessaloniki, Greece, the MSc degree in Radio Frequency Communication Systems from the University of Southampton, and the PhD degree in Electronic Engineering from Queen Mary University of London. Currently, he is a postdoctoral researcher with the Information Technologies Institute. His research interests include semantic multimedia analysis, indexing and information retrieval, data mining, search engines, and human interactions, as well security and environmental applications. Dr. Vrochidis has successfully participated in many European and national projects and he has been involved as a co-author in more than 55 related scientific journal, conference, and book chapter publications.