PESCaDO

FP7-IST-248594

Personalized Environmental Service Configuration and Delivery Orchestration

D8.1 Overview of the State of the Art

Due date of deliverable: 31.03.2010

Actual submission date: 31.03.2010

Start date of project: 1st January, 2010 Duration: 36 months

Lead contractor for this deliverable: FMI

D8.1	Overview of the State of the Art
Project Acronym :	PESCaDO
Contract No :	FP7 - 248594
Due Date :	31.03.2010
Reply To:	Ari Karppinen ari.karppinen@fmi.fi
Actual date of delivery:	31.03.2010

Deliverable Identification Sheet

Project ref. no.	FP7-IST-248594
Project acronym	PESCaDO
Project full title	Personalized Environmental Service Configuration and Delivery Orchestration
Security (distribution level)	PU
Contractual date of delivery	Month 3, 31.03.2010
Actual date of delivery	Month 3, 31.03.2010
Deliverable number	D8.1
Deliverable name	Overview of the State of the Art
Type	Report
Status & version	Submitted, v1
Number of pages	115
WP / Task responsible	FMI
Other contributors	All
Author(s)	All Partners of the Consortium
Internal Reviewer	Al partners of the Consortium were involved in the cross-reviewing of individual sections.
EC Project Officer	Manuel Monteiro
Abstract	The purpose of this deliverable is to evaluate the state-of-the art in the field of environmental service configuration and delivery in order to (i) decide on which grounds the R&D work in PESCaDO should be built; (ii) detect the shortcomings and restrictions of the current approaches and thus be able to contribute to their resolution. The deliverable thus discusses the state-of-the-art in environmental services, i.e., their coverage, purpose, targeted audience, information presentation techniques, etc., and assesses in detail the state of affairs in the individual research areas that are of relevance to PESCaDO. In the final wrap-up, the main challenges for PESCaDO in the light of the presented state-of-the-art are formulated.

1 INTRODUCTION

The purpose of this deliverable is to gain a detailed insight of the state of the art in technologies relevant to PESCaDO in order to be able to decide on which technologies PESCaDO can build upon and how these technologies must be improved for PESCaDO to achieve its objectives. In what follows, we start with the outline of a general view on environmental services, their orchestration and the motivation for the use of orchestrated services. Then, we familiarize the reader with the structure and content of the deliverable.

1.1 ENVIRONMENTAL SERVICES AND THEIR ORCHESTRATION: WHY AND FOR WHOM?

The role of meteorological, air quality, water quality, etc., in short, environmental, services in our society steadily increased over the years, reaching to date an unprecedented significance. This is, on the one hand, due to the increased sensitivity of the population for environmental issues, and, on the other hand, certainly also due to the advances in the quality of the services provided to the population.

In the present-day Europe, with its well-established national air quality and meteorological networks, there are solid ties between the air quality and meteorological agencies and seemingly well-connected data distribution networks. However, there is an increasing need for the orchestration of environmental services spread across the Web in order to provide users with personalized decision support or tailored environmental information. This is because: (i) as a rule, no single service provider covers the entire range of environmental conditions required for a comprehensive service; (ii) the delivery of external complementary data contracted by a service provider covers, as a rule, only the standard prominent parameters; (iii) the pressure for validation of own and external information to ensure a high quality service is increasingly high; (iv) the requirements of the users increasingly concern regions that are not covered by a provider, and the provider is not able to provide information or data on these regions; (v) the users are burdened with potentially contradicting and deviating information from several sources and with the assessment of the information for decision making. PESCaDO aims to meet this need for environmental service orchestration. PESCaDO will offer an image of an environmental service as an interconnected multipurpose user-oriented service for a federated community of citizens, public services (such as tourist offices and environmental institutions), public administrations, and entrepreneurs active in sectors sensitive to environmental conditions. A service, to which the user is able to communicate his/her needs, and which is able to provide decision support and deliver the necessary information in the language of user's preference.

Considerable contributions have already been made in many of the working areas that are involved in such an endeavor as PESCaDO. In the following sections, the most relevant of them are reviewed.

1.2 THE CONTENT AND STRUCTURE OF THIS DOCUMENT

The review starts with a global overview of existing environmental services (Section 2). In the following sections then the individual technologies, practices and tendencies are reviewed. The following themes are addressed in Sections 3 to 11:

(i) Discovery of environmental service nodes, where "node" is interpreted as any URI-carrying resource in the world wide web graph, and "service" is understood as provision of (meteorological, air quality, traffic-related, etc.) information considered useful for the targeted addressee(s). The provision can be realized in terms of the display of the information in a web page (for the consumption by a human reader) or in terms of a web service (for the consumption by a program).

(ii) Uncertainty metrics for the assessment of the trustworthiness and quality of the information provided by the individual environmental service nodes.

(iii) Technical protocols for connecting environmental services at different levels of the information and control flow.

(iv) Environmental node orchestration, where "orchestration" means the construction of a configuration of several environmental nodes such that: (a) their output data/information complement each other (as, e.g., the output of an air pollution prediction service and the output of a pollen prediction service); (b) their outputs compete with each other (as, e.g., the information provided by two different meteorological services for the same region); (c) the output of one service serves as input to another service (as, e.g., the output of a meteorological service is used by an air quality service).

(v) Ontology alignment and extension.

(vi) Distillation of content from multilingual environment material.

(vii) User-oriented reasoning and decision support strategies.

(viii) User-system interaction techniques.

(ix) Multilingual user-tailored environmental information synthesis and delivery.

After the discussion of the existing technologies in the PESCaDO-relevant areas, a number of conclusions for the development of the PESCaDO platform are drawn in Section 12.

2 A GLANCE AT EXISTING ENVIRONMENTAL SERVICES

This section is divided into three parts. In the first part, we give a short overview of the environmental services landscape. The second part addresses operational (public and commercial) environmental services, while the third part focuses on some most prominent prototypical R&D realizations of environmental services.

2.1 OVERVIEW OF ENVIRONMENTAL SERVICES

A great number of environmental services are available to date; most of them are web-based, although other communication channels such as TV, mobile phone services, (paper) press, and radio are used as well. The coverage, purpose and presentation of these services vary largely: among them are meteorological services, air quality services, water quality services, waste management services, leisure time services, and so on. Not all of them are primary from the viewpoint of service detection, orchestration and user-oriented information delivery – the topics that are central in PESCaDO. Therefore (and also due to the lack of space), we will have to restrict to a few services of some selected types of mainly web-based services. Anything else would be doomed to fail in the context of a project report.

Despite the multiplicity of the environmental services mentioned above, the majority of environmental web-based services still consists of meteorological services that offer basic information for general public. For bigger cities, a great number of competing services of this kind is available. Reported are, as a rule, temperature, wind, air pressure and humidity; the sunrise and sunset times may also be given. Some of these services (usually public services or services with a broader information spectrum) also offer more general background information on weather and related topics such as climate change (as we will point out later on, the offer of more general "educational" information related environmental issues is a distinctive characteristics of public services with a legal mandate). Consider, for illustration, the display of information in two meteorological services for Greater London in Figure 2.1 – one in Spanish and another one in English.

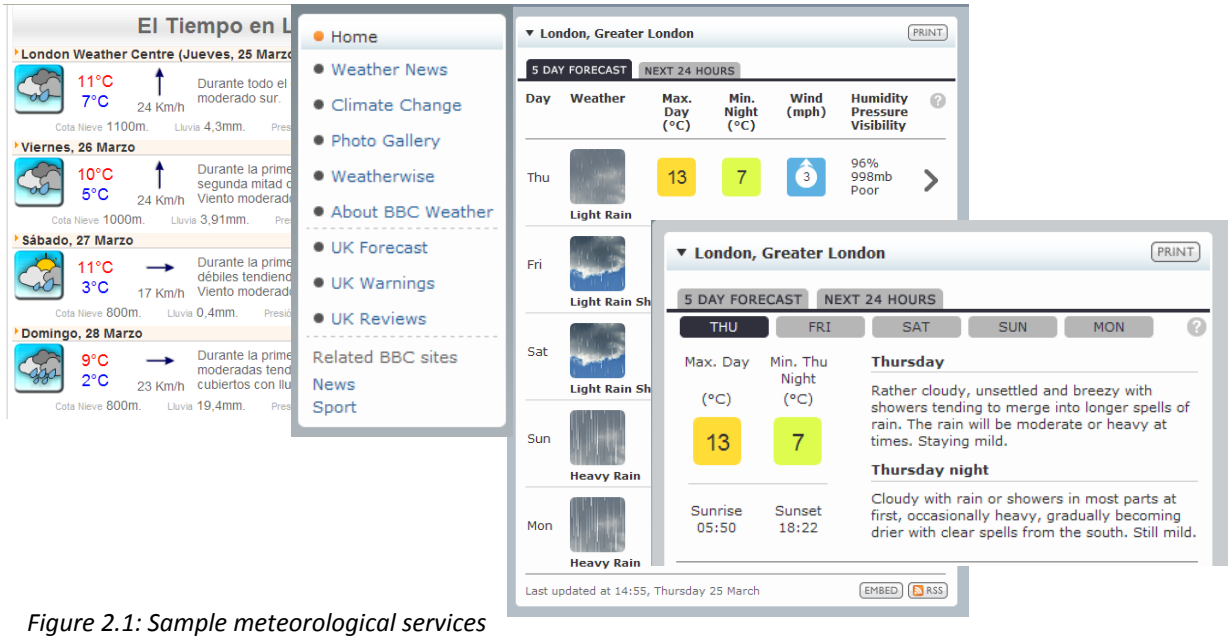


Figure 2.1: Sample meteorological services

Common are also meteorological services that target specific types of users – farmers, people pursuing certain kinds of outdoor leisure activities (such as hiking, mountain climbing, sailing, and the like), etc. This is especially the case in regions with abrupt weather changes and for users who depend on the weather; cf., e.g., a sample service for cross-country skiing (in French) and for sailing in Figure 2.2.

Meteorological services are often completed by traffic services (concerning so-called "road weather") – especially in regions where the weather has a strong influence on traffic conditions. The inclusion of traffic information links may also be restricted to a specific season (usually winter, but can also be, e.g., autumn in the case of storms); cf. Figure 2.3 below.

Increasingly common are air quality services that may be combined with conventional meteorological services (as, e.g., in the case of many newspaper web-pages) or offered as a separate service (as in the case of FMI (http://www.fmi.fi/products/air.html), HSY (http://www.hsy.fi/), etc.). The big difference between the meteorological and related services and air quality services is that a great number of the former is provided by private companies who either fully specialize on them or who offer them as a kind of add-on – as, e.g., the on-line press. In contrast, most air quality services are provided by governmental or public institutions that implement national or provincial legislation – although critical air quality information (e.g., on concentrations of allergy causing pollen concentrations during the relevant seasons) is also provided by commercial mainly meteorological services. These institutions have thus the mandate to inform and also to educate the population on the topic of air quality. This is usually reflected in the range of information being offered: the information contains not only measured or forecasted air quality data, but also background information, links to health issues that are related to air pollution, etc. (see, however, Figure 2.1, where background information and links to further related information is offered in the context of a meteorological service). The FMI AQ-service is a good example for such services.

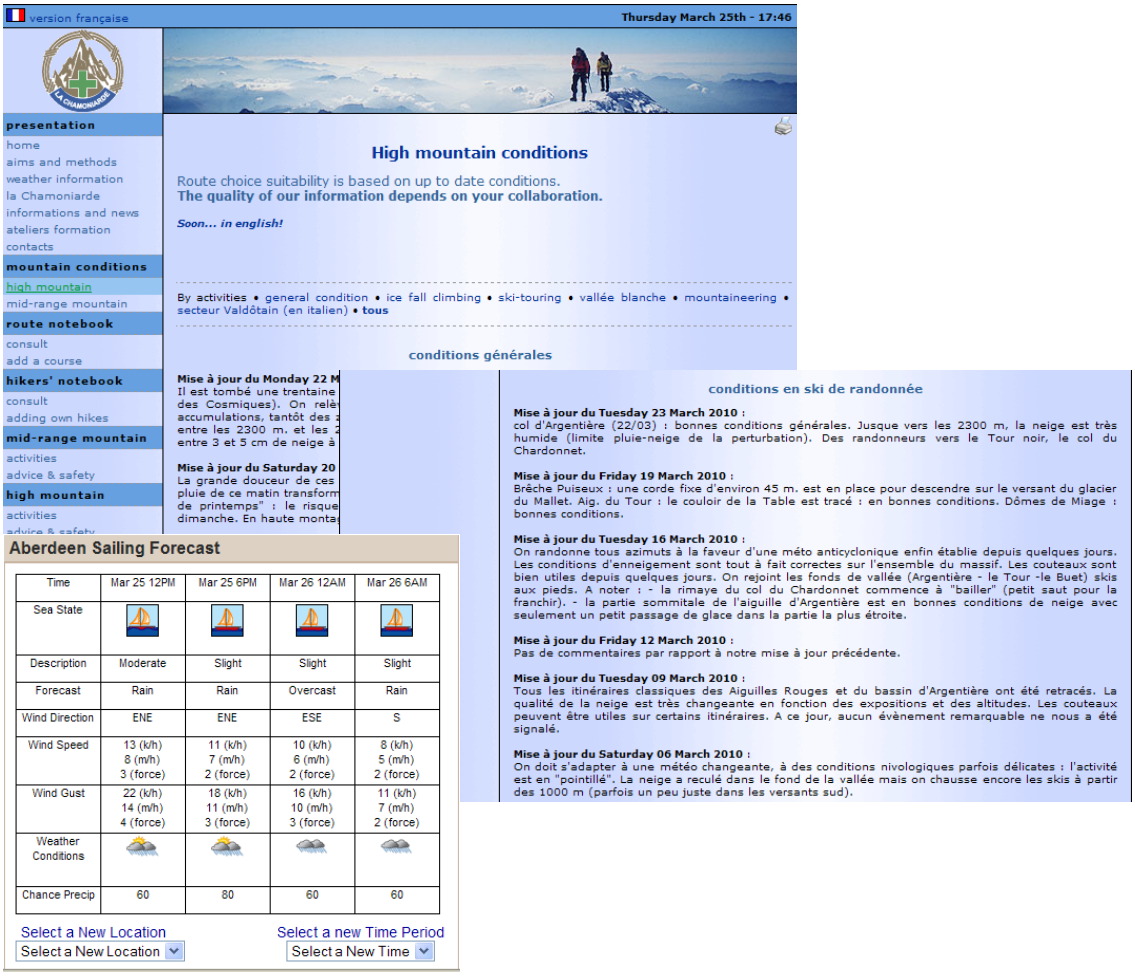


Figure 2.2: Service targeting a specific group of addressees (here: cross-country skiing and sailing)

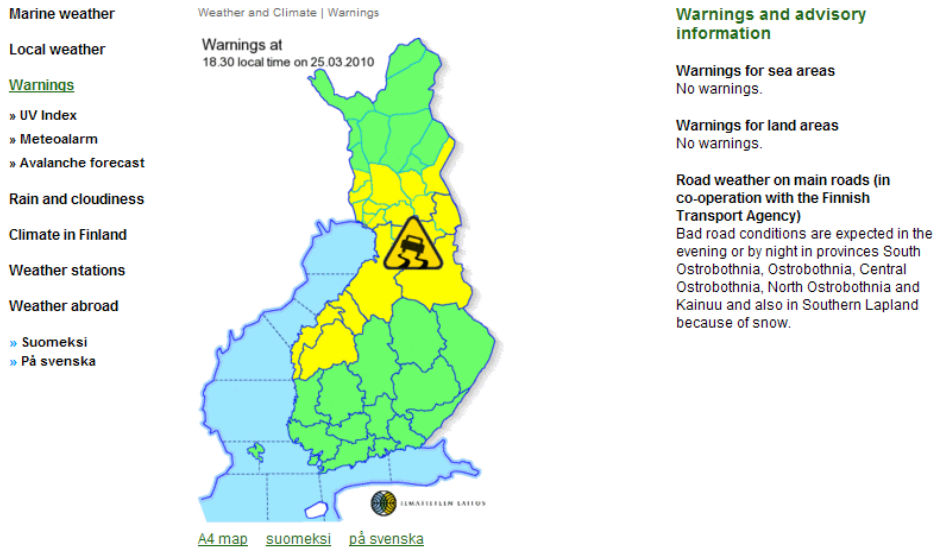


Figure 2.3: Integration of "road weather" into a meteorological service

A few (provincial or central) government-supported or municipal services (such as, e.g., the LUBW-service http://www.lubw.baden-wuerttemberg.de); cf., also Subsection 2.2.3 below) cover all major aspects of environmental information management and communication – including, among others, air quality, climate information, waste

management, noise pollution, and water management). To be noticed is also – even if this is not visible for the end user who accesses the information via the webpage of the service – that such services may be multipurpose services. For instance, the LUBW-service offers an intranet subservice that is supposed to fulfill the needs of the administrative staff of the municipalities in the Land Baden-Württemberg. However, such comprehensive services are still rather uncommon.

With respect to user-orientation of the environmental services that are available to date, it is to be noted that this relevant feature has not yet been taken into account to a sufficient degree. As mentioned above (cf. also Figure 2.2), some operational, first of all meteorological, services may well provide user-specific information (mainly for users who practice different kinds of sports). However, a personalization of the information as targeted by PESCaDO does not take place. The only exception we are aware of is the prototypical air quality service MARQUIS (cf. Subsection 2.3.1), which has a fine-grained user profile typology that can be further personalized by any user.

As far as the modi used to communicate environmental information are concerned: pictograms are the most often used mode; often, also simple text statements are provided. These statements are, as a rule, written manually – again, with the exception of a few prototypical services; cf. Section 2.3.

In what follows, we consider some selected operational environmental information systems (as run, for instance, by public or governmental institutions) and prototypical services that emerged as results of R&D projects, but, in their majority, did not reach a continuous operational state. The first demonstrate what kind of services an end user is offered to date, while the latter show what kind of techniques is currently being worked on.

2.2 Operational environmental services

In this section, we consider in some detail a few different environmental services that are of particular importance to PESCaDO. These are, first, the services offered by the PESCaDO partners FMI (2.2.1) and HSY (2.2.2) and by the PESCaDO affiliated user LUBW (2.2.3). In Subsection 2.2.4, we then take a glance at several other prominent environmental (first of all, air quality) services.

2.2.1 Example of a centralized national environmental service – case Finland

In what follows, we discuss two of the FMI services: the air quality service and the meteorological and environmental testbed service.

2.2.1.1 FMI air quality service

In Finland, it is enacted in the legislation that the municipalities are responsible for arranging sufficient air quality monitoring in their own territory. Due to this, there are currently some 36 different air quality monitoring networks in Finland. The municipalities are also required to inform and advise the population if and when different limit or threshold values are exceeded, which in the past resulted in very diverse information systems, ranging from annual printed summary reports to real-time web based services, depending on the technical and financial resources of the individual actors. This set the population in a very unequal position with respect to getting information about the air quality in their own surroundings. Therefore a centralized web service *www.ilmanlaatu.fi* (*www.airquality.fi*, *www.luftkvalitet.fi*) was built to collect the air quality observations from all monitoring networks in one data base and disseminate them on a public web site in real time. The building of the service was initiated by the Finnish Ministry of the Environment, together with the Finnish Meteorological Institute who also carried out the actual planning and implementation of the service.

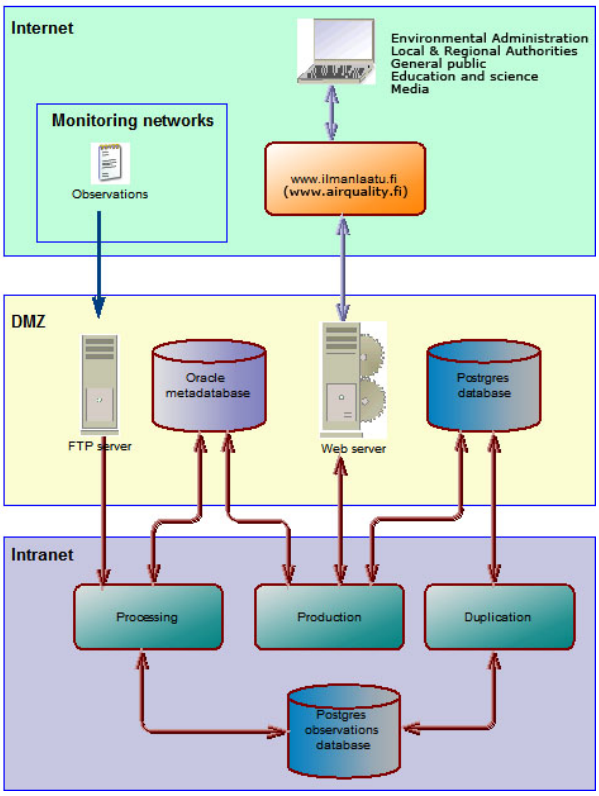


Figure 2.4: The architecture of the Finnish national air quality portal.

The air quality portal is maintained by the Finnish Meteorological Institute and offers its services free of charge to the municipalities in Finland as a convenient way of publishing their air quality observations in real time in a harmonized way. The portal also serves as the national air quality data repository, because, after validation, the data is kept in the history data base where it can be downloaded (in csv or ssv format) for e.g. reporting and research purposes. A simplified diagram of the portal architecture is presented in Figure 2.4 above. The monitoring networks automatically upload their data once an hour to the portal server where it goes through routine checking for missing, suspect or rejectable values as well as air quality index and statistics calculation, and is subsequently stored in the portal data bases. Not shown in Figure 2.4 is the web interface built for the monitoring networks to maintain the metadata of their own measurements directly in the portal data base and perform quality control on their observations.

In addition to helping the municipalities to publish their real time air quality observations and store the historical air quality data, the portal is designed to enhance environmental awareness, and to serve the general population in decision making in their everyday life. That the portal is also used for this purpose is especially evident in the springtime when episodes of high street dust concentrations (manifesting as high PM10 concentrations) are frequently observed in almost all major cities in Finland: the portal typically receives one third of its annual visitors in the months March to May. Other important user groups for the real time air quality information are the environmental administration and the local and regional authorities. The main language of the portal is Finnish, but the core content, i.e. the prevailing air quality situation and the real time observations, is also available in Swedish and English.

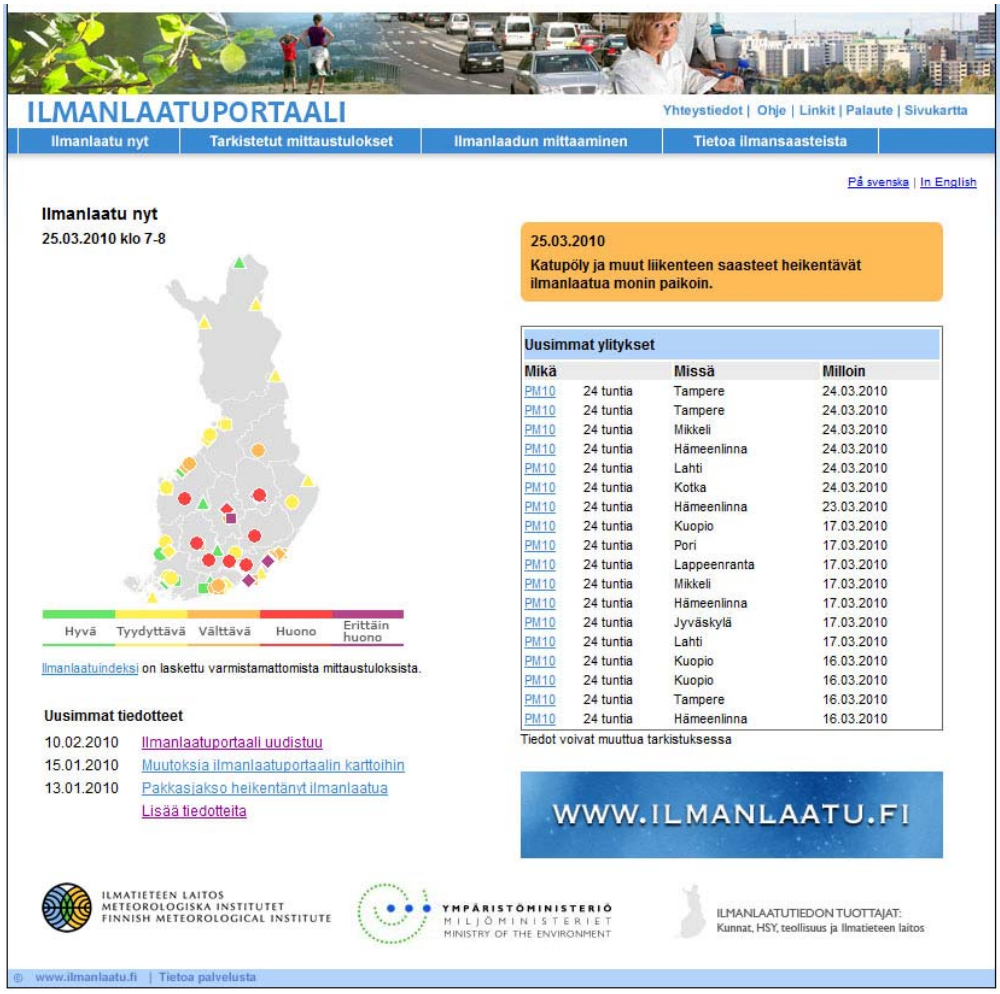


Figure 2.5: The front page of the Finnish national air quality portal with the map presentation of the prevailing air quality index, recent bulletins, list of the latest exceedances and the orange alert box.

The front page of the service (Figure 2.5) gives an at-a-glance overview of the prevailing air quality in Finland, represented by the Finnish air quality index (see section 2.2.2.1 for more information) at the measuring sites. It also shows lists of recent exceedances ("Uusimmat ylitykset") of limit or threshold values and air quality bulletins ("Uusimmat tiedotteet"). In exceptional air quality situations, an orange alert box with specifics is placed on the front page. In the typical springtime episodic situation in Figure 2.5 the bulletin tells that "street dust and other traffic originated pollutants may cause degradation of air quality in several locations".

In the Air Quality Now –section of the portal the visitors can make more detailed selections with respect to what regions or locations and which pollutants they want to view (Figure 2.6). They can also select how they want the data to be presented (map, graph or table), the scales of the graphs. For some pollutants they can also select certain relevant statistics (e.g. 24 hour averages for PM10). The air quality index can only be viewed in map or graph presentations while the table choice is available for all measured quantities. This restriction with the air quality index is because even in Finland there may be several different ways of calculating the index, all yielding different numerical values even though the triggering pollutant and the resulting classification of the air quality remain exactly the same.

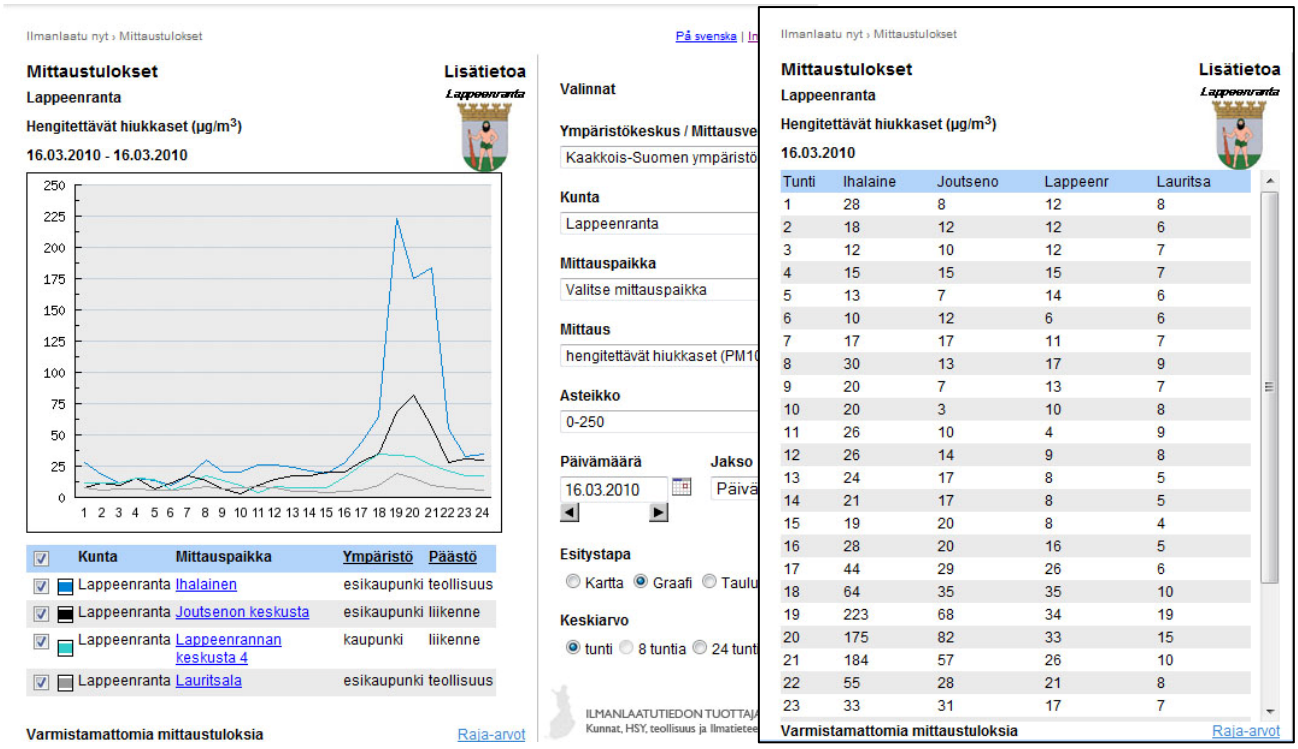


Figure 2.6: Examples of graphical and table presentations of hourly PM10 concentrations in the municipality of Lappeenranta in south-east Finland during a street dust episode in March 2010. Next to the graph there are several drop-down lists for choosing the region, measuring site, the quantity to be viewed, y-axis scale, date, the displayed period (day/week/month are available) and the averaging time. In the table only one day at a time can be displayed to discourage large scale downloading and use of real time data which is subject to change any time.

Forecasted air quality is not yet available in the portal, but as a first step in this direction, the predicted dispersion of smoke (i.e. PM2.5 concentration) from wildfires for the next 18 hours is presented as an animated map (Figure 2.7).

In addition to the real time observations, and the historical air quality data, the national air quality portal gives general information on air pollutants and their health and environmental effects as well as air on quality management and measurement techniques. It also provides detailed descriptions of the Finnish air quality measuring stations and their instrumental repertoire. Furthermore, the portal contains information on the health effects of air pollutants and gives advice on how to reduce your exposure to pollutants and relieve possible symptoms in episodic situations (Figure 2.8). The portal also gives the visitors practical advice on what they themselves can do in order to improve the air quality in their immediate surroundings. To benefit the visitors seeking more profound knowledge of air quality issues, there are advanced articles on specific subjects, written by prominent experts in the field, as well as links to sites with even more detailed information.

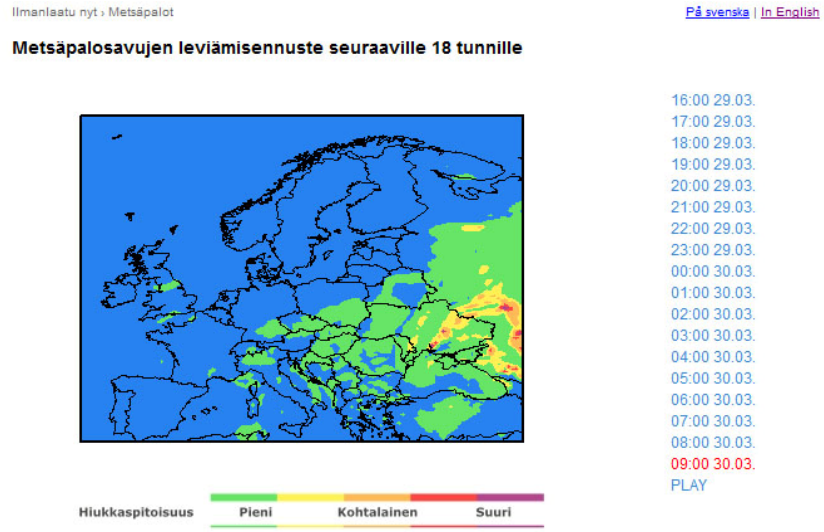


Figure 2.7: The predicted concentration of fine particles (PM2.5) from forest fires. The concentration is low in green areas, moderate in yellow and orange areas, and high in red areas. The forecast is calculated by the SILAM air quality model of the Finnish Meteorological Institute.

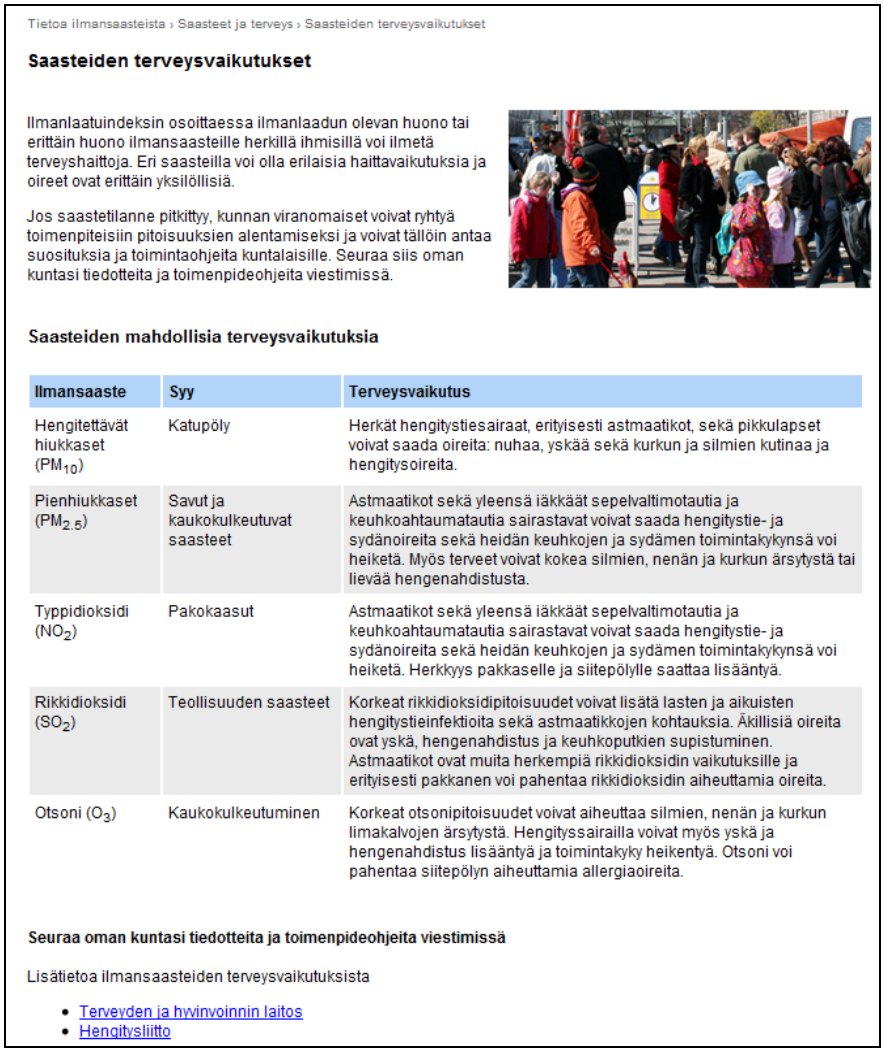


Figure 2.8: The portal gives information about the health effects of air pollutants.

2.2.1.2 HELSINKI TESTBED: METEOROLOGICAL AND ENVIRONMENTAL SERVICE

The Helsinki Testbed (HTB) [1] was originally a research program; nowadays also an operational meteorological and environmental service designed to provide both on-line and off-line information on mesoscale weather phenomena, urban and regional forecast and dispersion modeling and their verification, applications in a high-latitude coastal environment, and data distribution for the public and the research community.

The domain of the Helsinki Testbed observing network, roughly 150 km x 150 km, covers much of southern Finland and the Gulf of Finland and includes Finland's most populous city, Helsinki (Figure 2.9). The Helsinki Testbed is composed of a variety of different observing instruments. In addition to 46 existing weather stations operated by FMI, an instrumented 320m tall radio mast, and a 145m tall mast at a nuclear power plant, originally more than 100 new weather transmitters were added, including more than 40 communication mast sites of 60–100m height. The weather transmitters measure temperature, humidity, air pressure, rain, and wind speed and direction. Other data sources have included e.g radiosonde data, road weather observations, a hydrometeor size detector, total lightning location system, a Doppler lidar, and a Doppler sodar.

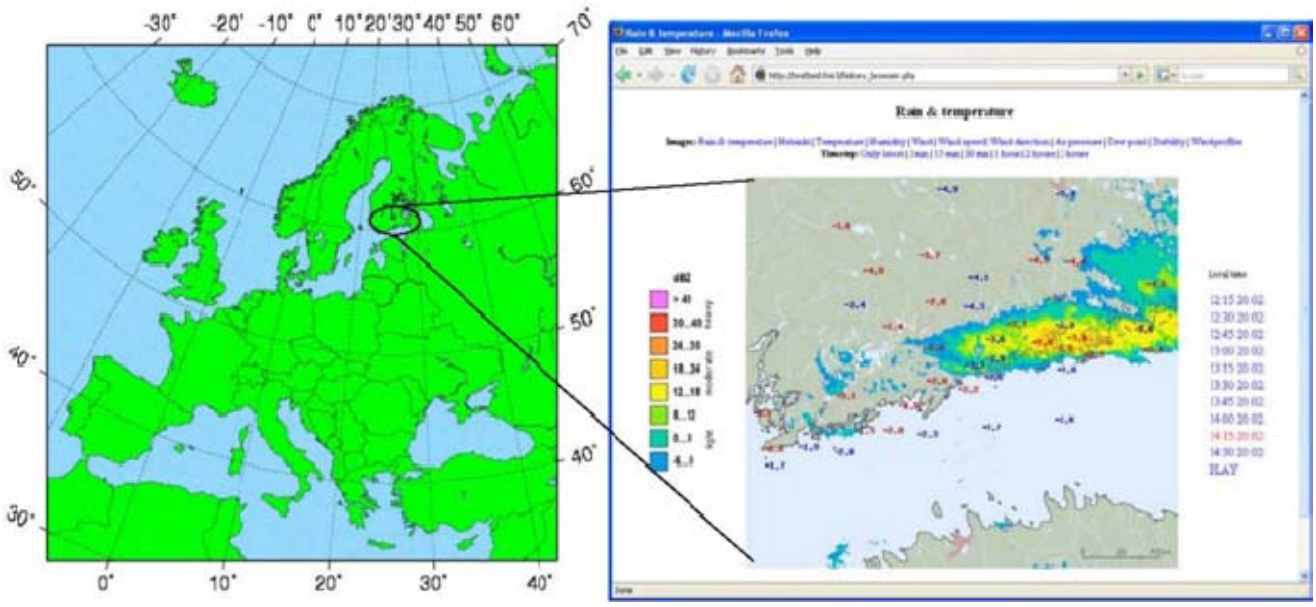


Figure 2.9: The approximate location of the Helsinki Testbed domain (left), and the most popular animation page of the http://testbed.fmi.fi Web site (right).

HTB-measurements include also the University of Helsinki's and FMI's SMEAR-III [2]urban measuring station consisting of a 31m tower equipped with meteorological instrumentation at several heights. Measurements include profiles of the temperature and wind and radiation components (Figure 2.10). The fluxes of sensible heat, momentum, carbon dioxide and water vapor are measured by the eddy-covariance technique. Next to the tower is situated an air-conditioned container where a diversity of aerosol particle and gas concentration instrumentation is located.

The current focus in the development of the TestBed platform has shifted from just collecting observations to developing services benefiting citizens, businesses, governmental authorities, and industries. Numerous pilot services

[1] *http://testbed.fmi.fi*

[2] http://www.atm.helsinki.fi/SMEAR/index.php?option=com_content&task=view&id=25&Itemid=59

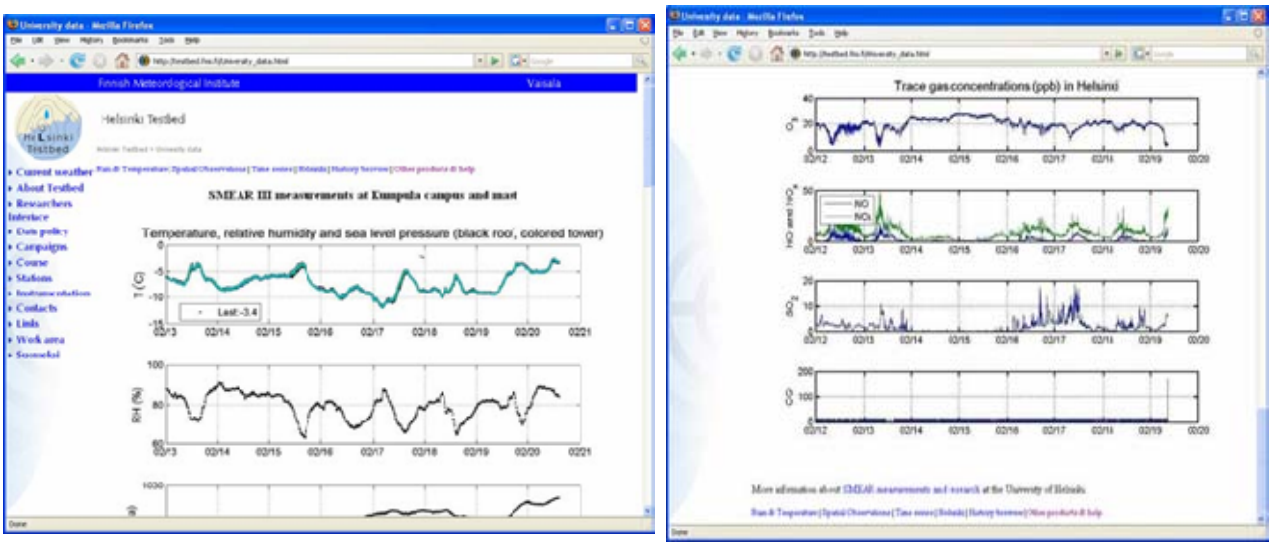


Figure 2.10: SMEAR-III measurements on Testbed Web site.

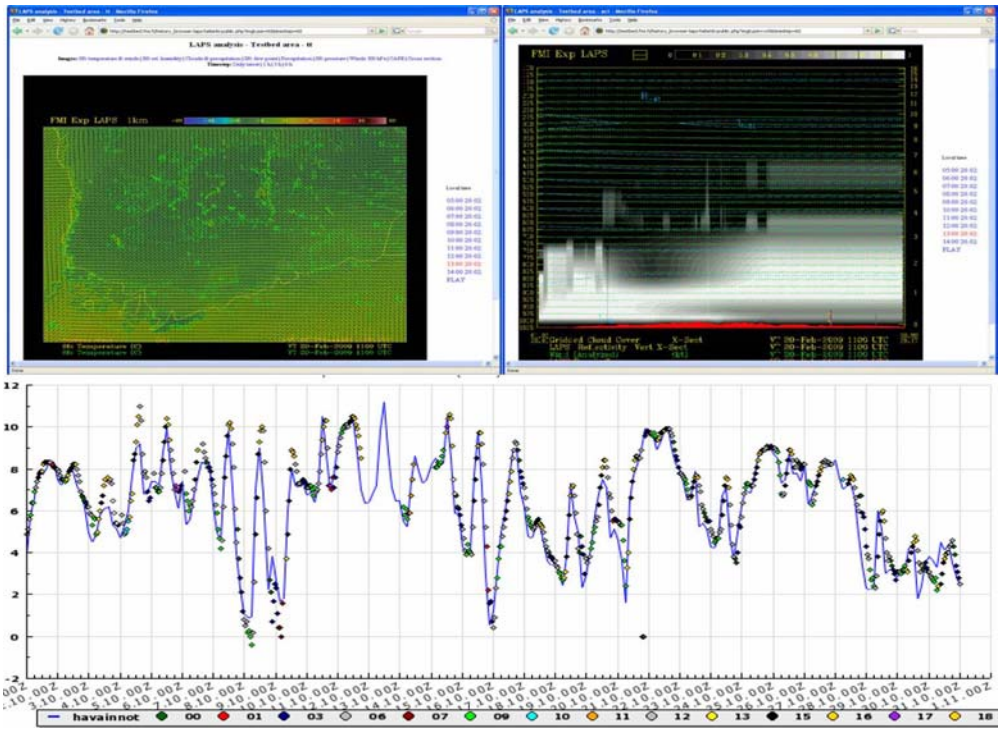


Figure 2.11: Upper panel: LAPS temperature analysis for the Southern Finland area (left), and analysis east-west cross-section at the middle of the domain (right). Lower panel: LAPS surface analyses (colored dots) compared with an independent road weather station temperature observation (ºC, blue line) in the city of Nokia, Finland. The data is shown for October 2008.

for end users have been developed for weather, air quality and traffic weather services. One example of the new products developed based on the utilization of the "The Local Analysis and Prediction System" (LAPS)[3] is illustrated in Figure 2.11.

The HTB portal includes an internet service ("Researcher's Interface ") – see Figure 2.12 – open for anybody interested in meteorological and environmental data from the area. It enables users to make direct HTB-database queries. Users must register to the service by logging on to the authentication service including acceptance of the terms and conditions of the User Agreement and the Appendices. Queries of online and off-line data, station information (meta-data) and plots and maps generated from the data can be performed with interactive drop-off boxes and menus or by a direct service request to the database using a free-form XML-request. Some search templates and documentation of the specific XML-schema used for the queries are available at the web-portal.

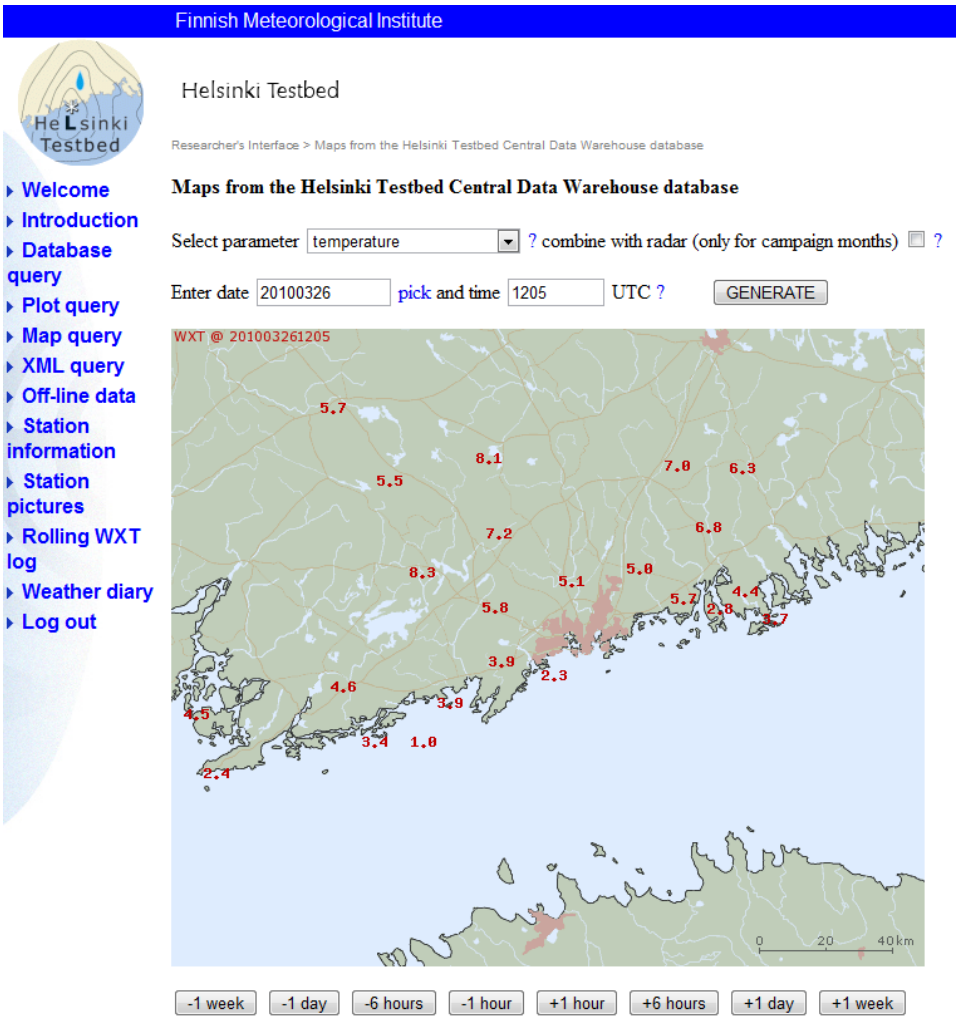


Figure 2.12: Testbed Researcher Interface.

[3] http://laps.noaa.gov/

2.2.2 THE SERVICES FOR THE HELSINKI METROPOLITAN AREA

2.2.2.1 THE HSY AIR QUALITY SERVICE

The HSY air quality web service *http://www.hsy.fi/seututieto/ilmanlaatu* provides air quality information in real time for the Helsinki metropolitan area. The front page of the service (Figure 2.13) provides an at a glance view of the current air quality in the region, according to the Finnish air quality index. The service is designed to guide the population in decision making in their everyday life. The main elements are the map presentation of the regional air quality according to the national air quality index, and the daily description of the air quality situation ("Tilannekuvaus arkipäivisin") which is available on weekdays. All daily descriptions are accessible in text form in the web address

Figure 2.13: The front page of the HSY air quality web service. The different symbols on the map represent measurements made in different environments, ranging from traffic and street canyons (cars and cars between tall buildings) to residential (houses and landscape) or background (trees) areas, with the colour representing the air quality classification through the five classes from good to very poor.

http://ilmansuojelu.ytv.kaapeli.fi/info/tilannekuvaukset_alku.php. Also a list of the latest air quality bulletins ("Ajankohtaista") is offered to the public. Air quality in the Helsinki metropolitan region can also be followed by a mobile service (http://mobi.hsy.fi/ilmanlaatu).

The Finnish air quality index has been developed by the HSY together with national health experts, such as the National Institute for Health and Welfare (THL). The Finnish air quality index differs from the indices used in most other European countries, with much lower concentration limits for classifying the air quality through the five classes from good to very poor. This is because the air quality in Finland, due to the remote location of the country and the relatively sparse population and thus low emissions, is, in general, better than e.g. in the densely populated Central Europe. If an air quality index tailored for the more polluted regions of Europe would be applied to Finland, it would most probably always show good or very good (i.e. "green") air quality and the information value of the index in guiding the population in decision making (the main purpose of developing the indices) in their everyday life would be nil.

However, Finland does have some specific air quality problems, typical also in other North European countries, where the national index is very important. Among these, the wintertime inversions, when pollution from traffic (especially NOx) is trapped in the very shallow surface layer where the concentrations can increase to harmful levels, springtime street dust episodes with their extremely high PM10 concentrations, and the spring/early summer ozone episodes, fed by long range transport and intensified by domestic emissions in sunny high pressure conditions typical of the season, are the most prominent.

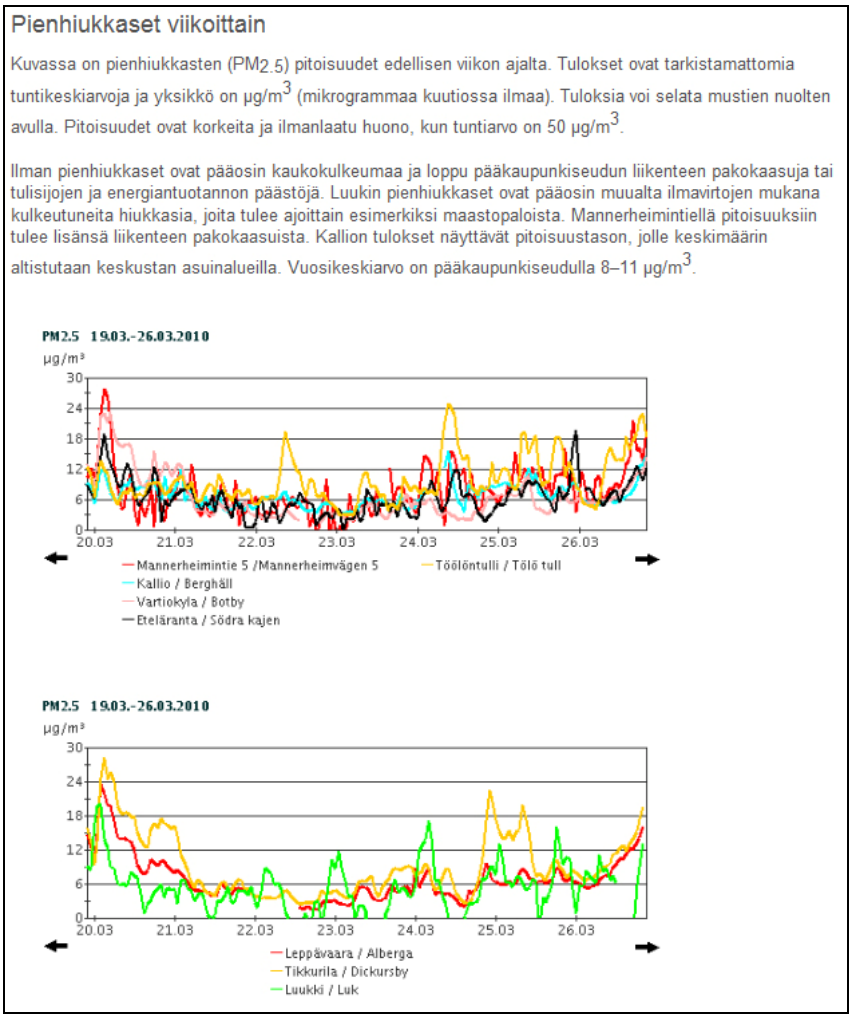
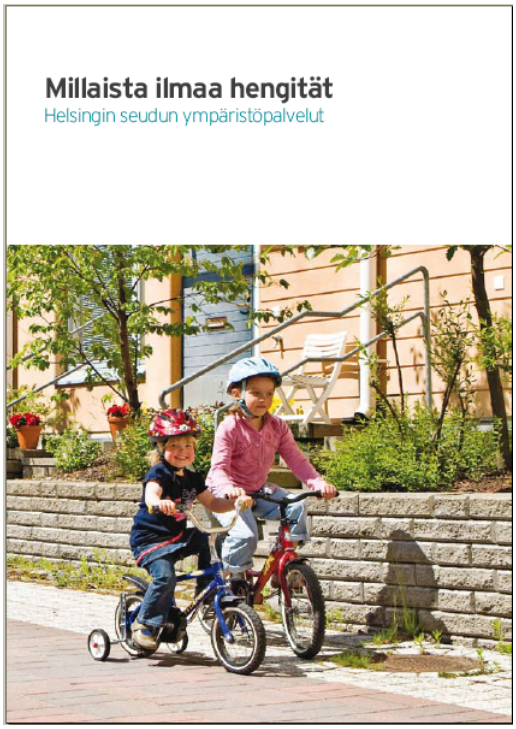


Figure 2.14: The concentrations of fine particles (PM2.5) can be viewed as 24 hour or weekly graphs.

In addition to the map presentation, the concentrations measured at each of the monitoring stations (currently eleven) can be viewed as 24 hour or weekly graphs or by location (example in Figure 2.14). No download of the data is available, but the validated monitoring results from previous years are accessible through the national Air Quality Portal described above, similar to all other Finnish air quality data.

Besides the web service, HSY is also directly connected with the media, and information about current air quality in the Helsinki metropolitan region is broadcast in the TV and radio, as well as the leading national newspaper every morning on weekdays. There are also several (currently eight) air quality screens in various locations around the metropolitan region so that the public can be kept up to date of the latest developments in air quality.

HSY publishes annual reports of the air quality in the Helsinki metropolitan region and popular leaflets on air quality related topics (Figure 2.15), all available also through their web service, in order to communicate the latest developments and advance the general awareness of the population of air quality issues.



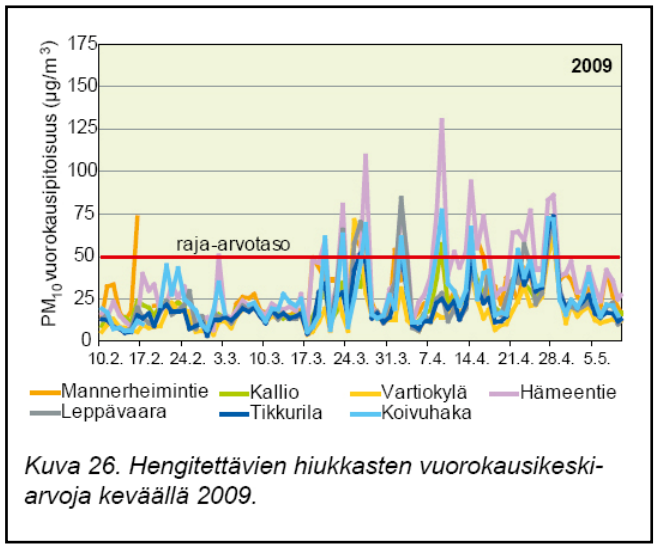


Figure 2.15: An example of a popular air quality pamphlet published by HSY in 2010. The 16 page pamphlet was titled, and also answered, the topical question of "What kind of air do you breathe". To the right is an example from an annual air quality report, showing the daily averages of respirable particles (PM10) in the spring of 2009.

2.2.2.2 THE TRAFFIC SERVICE IN THE HELSINKI METROPOLITAN AREA

The Destia traffic service (*http://www.destiatraffic.fi/liikenne*) provides information about the traffic situation in Finland and can be zoomed in the browser to different regions (Figure 2.16). The service offers information about traffic jams, bulletins, road weather and web cameras, the travelling time and road works. The traffic jams are presented in colour (yellow = slow, red = jammed) and additional information about the traffic jams is provided in a pop-up window.

The HSL/HRT (Helsingin seudun liikenne / Helsinki Region Transport) service (*http://www.liikenteeseen.fi*) is the traffic information portal of the Helsinki metropolitan area. It combines information from several sources, in order to help plan and compare between various modes of travel. The service combines the regional traffic situation, traffic weather, route search, air quality etc. (Figure 2.17). It displays the traffic disturbances in a detailed way. The route

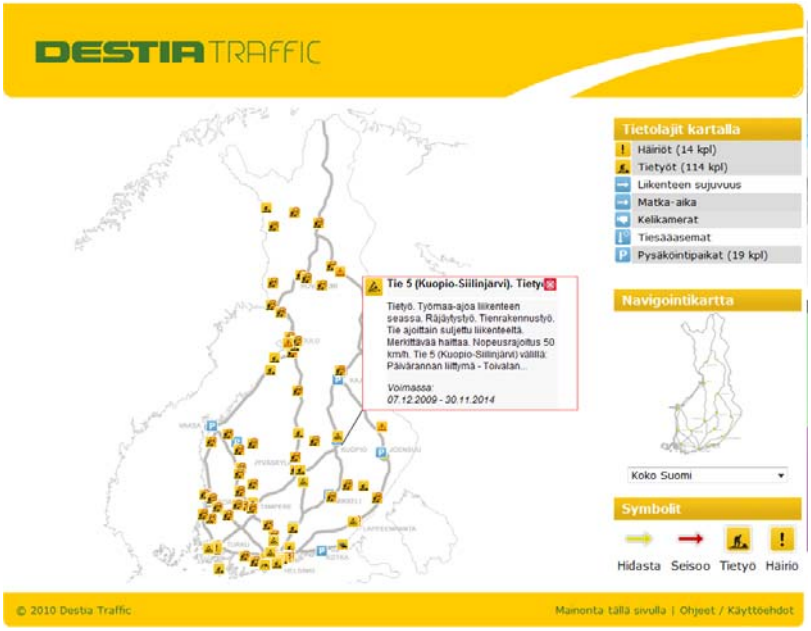


Figure 2.16: Traffic information on the front page of the Destia traffic service in Finland.

search finds the quickest route by public transport in the region. The service is available in Finnish, Swedish and English.

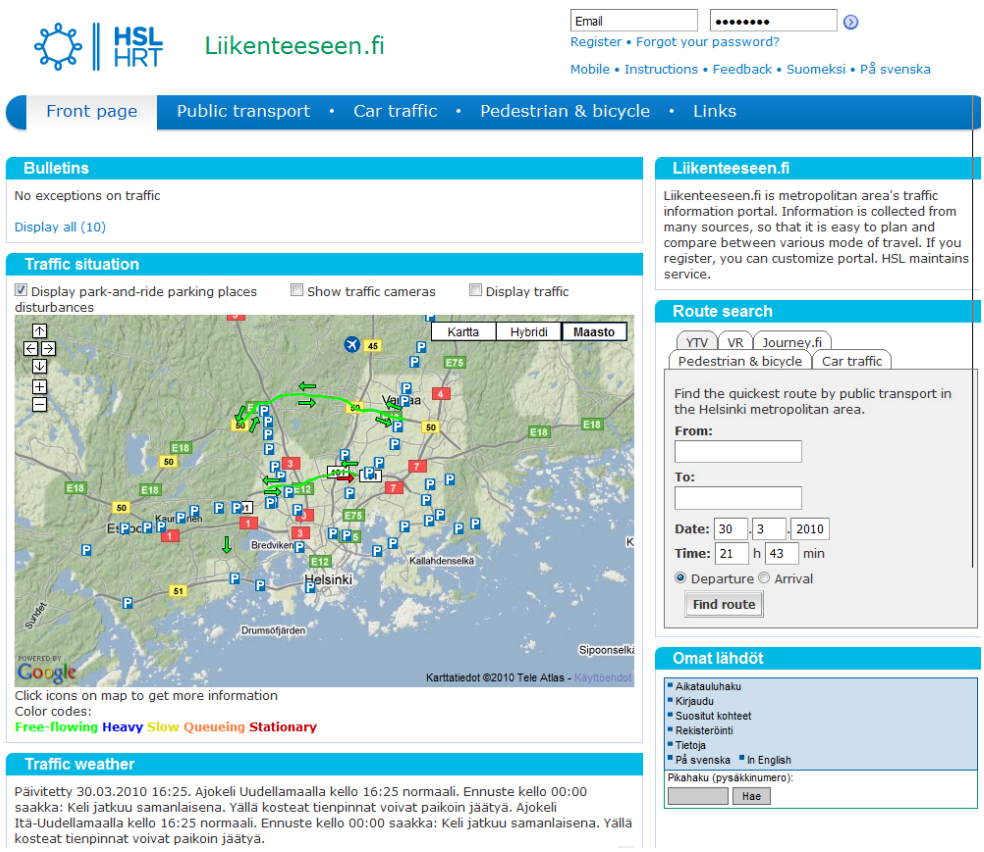


Figure 2.17. The front page of the HSL service Liikenteeseen.fi. The traffic situation is presented on the map and with the colour representing the traffic situation from free flowing to stationary.

2.2.3 The Environmental Information System of Baden-Württemberg

The Environmental Information System[4] of the federal state of Baden-Württemberg in Southwest Germany is the organizational, technical and content-related framework for the processing of environmental information within the state and municipal administration in Baden-Württemberg. It also provides the basis for reporting to the public. It is part of the e-Government Concept Baden Württemberg as well as of the ICT cooperation between the state and the municipalities on the one hand and the state, the Federation and the European Union on the other hand. The most important aspects w.r.t. PESCaDO of the Environmental Information System of Baden-Württemberg (UIS BW) are outlined in this subsection.

2.2.3.1 Architectural Outline

The UIS BW holds central data in three major databases; cf. Figure 2.18:

(i) The UIS reference database for the data exchange between the administrations,

(ii) The RIPS Pool which merges spatial data from various sources, and

(iii) The Measuring Series Operation System (MEROS) providing a uniform modeling for all measuring data from state-wide monitoring networks, covering various areas like water, soil, air and radioactivity.

These databases have been combined by ER modeling to one logical database, the Database for the Comprehensive UIS Components. All application data and spatial data of this encompassing reporting data base can be accessed by the UIS Reporting System, the component which finally generates the reports.

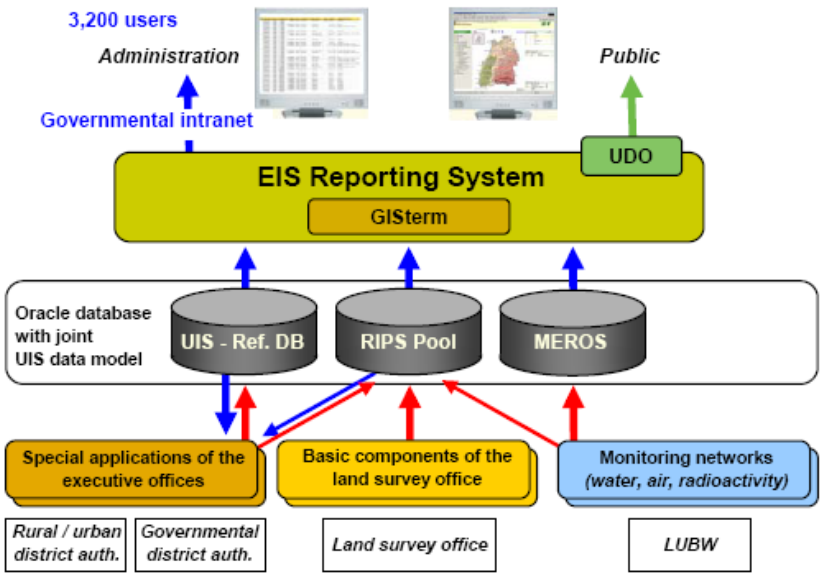


Figure 2.18: The architecture of the Environmental Information System Baden-Württemberg

[4] Adree Keitel, Roland Mayer-Föll, Albrecht Schultze: Framework Conception for the Environmental Information System of Baden-Württemberg. Proceedings of the European Conference TOWARDS eENVIRONMENT, March 25-27, 2009, pages 461-468. Prague, Czech Republic.

According to the status and the special needs of the users working with the UIS Reporting System, three different versions of the system have been deployed:

- A full version for specialists in the environmental administration, who need specific data analysis.
- A less complex web version (BRSWeb) for occasional users in the administration without special knowledge, who want to access and visualise data quickly over the administrative networks.

A particular version of the BRSWeb, called Environmental Databases and Maps online (UDO)[5], which offers public access to a limited amount of application data and spatial data over the internet. This version especially helps to fulfil the demands of the Environmental Information Act of Baden-Württemberg, which implements the demands of the directive 2003/4/EC of the European Parliament and of the Council on public access to environmental information on the state level.

2.2.3.2 Public Access to Environmental Databases and Maps

The interactive service UDO enables public access to selected environmental data stemming from governmental measurement and evaluation programs and to resources of digital maps. A number of predefined queries for the construction of reports are available. The themes, for which data and maps are available, are air quality, noise, radio activity, climate and regenerative energies, environmental meteorology, waste, water and specific spatial data. The following example (Figure 2.19) illustrates the retrieval of a series of air pollutants:

The query can be constructed interactively by selection of station, component, aggregation/frequency and period:

Figure 2.19: Selection of air pollutants for retrieval

The constructed query can be displayed before it is submitted and the results can be displayed directly in a web page; cf. Figure 2.20:

[5] http://brsweb.lubw.baden-wuerttemberg.de/

Luft/Stationsvergleich

Seite 1 / 1

Kriterium	logischer Operator	relationaler Operator	Wert
Station		=	Karlsruhe Mitte KA-M
Komponente		=	Ozon O3 µg/m³
Aggregation / Periode		=	Stundenwert fest
Zeitraum		=	23.02.2008 12:00 23.02.2008 18:00

Seite 1 / 1

Luft/Stationsvergleich

Seite 1 / 1

Stationsnummer	Station	Komponente	Datum / Uhrzeit	Wert	Einheit	Aggregationszeitraum	Aggregationsart
4441	Karlsruhe Mitte	Ozon	2008-02-23 13:00	61	µg/m³	Stundenwert	Einzelwert
4441	Karlsruhe Mitte	Ozon	2008-02-23 14:00	67	µg/m³	Stundenwert	Einzelwert
4441	Karlsruhe Mitte	Ozon	2008-02-23 15:00	70	µg/m³	Stundenwert	Einzelwert
4441	Karlsruhe Mitte	Ozon	2008-02-23 16:00	71	µg/m³	Stundenwert	Einzelwert
4441	Karlsruhe Mitte	Ozon	2008-02-23 17:00	70	µg/m³	Stundenwert	Einzelwert
4441	Karlsruhe Mitte	Ozon	2008-02-23 18:00	54	µg/m³	Stundenwert	Einzelwert
4441	Karlsruhe Mitte	Ozon	2008-02-23 19:00	28	µg/m³	Stundenwert	Einzelwert

Seite 1 / 1

Figure 2.20: Display of the query before submission

Nördlicher Oberrhein Ozon in µg/m^3 23.02.2010 15:00

Ozon	23.02.2010			22.02.2010	
Messstation	Aktuelles 1-Std. Mittel	Tagesmax. 1-Std. Mittel	Tagesmax. 8-Std. Mittel gleitend	Tagesmax. 1-Std. Mittel	Tagesmax. 8-Std. Mittel gleitend
Mannheim-Nord	38	51	43	58	43
Mannheim-Mitte	56	58	50	65	42
Odenwald	68	70	68	73	67
Mannheim-Süd	60	60	52	70	49
Heidelberg	52	57	49	55	58
Wiesloch	67	70	59	65	54
Eggenstein	66	77	69	83	57
Karlsruhe-Nordwest	50	71	64	70	44
Karlsruhe-Mitte	55	71	61	68	40
Schwellenwerte	180	180	120	180	120

Figure 2.21: Display of the retrieved data.

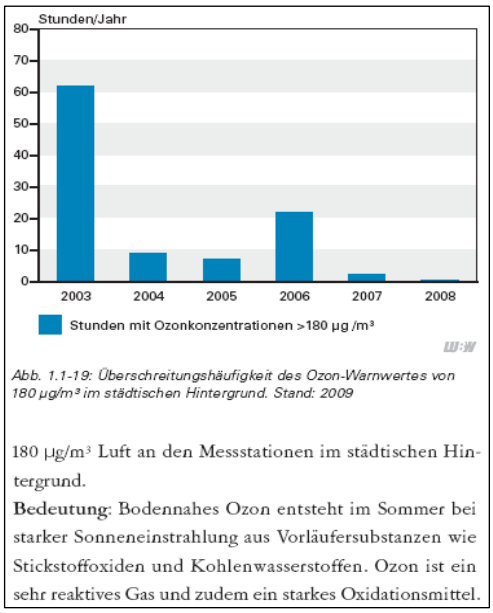


Abb. 1.1-19: Überschreitungshäufigkeit des Ozon-Warnwertes von 180 µg/m³ im städtischen Hintergrund. Stand: 2009

180 µg/m³ Luft an den Messstationen im städtischen Hintergrund.

Bedeutung: Bodennahes Ozon entsteht im Sommer bei starker Sonneneinstrahlung aus Vorläufersubstanzen wie Stickstoffoxiden und Kohlenwasserstoffen. Ozon ist ein sehr reaktives Gas und zudem ein starkes Oxidationsmittel.

Figure 2.22: Excerpt of a report

Figure 2.22 shows an excerpt of the 2009 report. The chart generated by the reporting system shows the exceedance of critical ozone values in the recent years.

The retrieved data can be presented in various formats and further processed in various ways, e.g. HTML, pre-assembled report in PDF or Excel formats, charts or email forwarding in defined cases (e.g. limit exceedance); cf. Figure 2.22 for a comprehensive report "Environmental Data in Baden-Württemberg", which reports on the entire environmental situation, is established once a year and published on the UIS BW portal in a PDF file[6].

All charts are accompanied with textual information, e.g. basic explanations or trend observations.

The portal also offers predefined queries for the retrieving information of high actuality which is not always possible by using conventional search engines . The following figure shows retrieved current ozone values of the region "northern upper-rhine" (current 1 hour average value, daily maximum based on 1 rsp. 8 hour average values).

2.2.3.3 PORTALS: NAVIGATION AND SEARCH FACILITIES FOR DATA AND TEXT

The illustration so far has concentrated on the provision of data in form of reports. In addition to this, a requirement for reporting information to the public is to offer a transparent presentation with a combination of reports, documents, catalogue information like metadata, databases and news. The prime example for an environmental information portal is the German Environmental Information Portal (PortalU)[7] which is developed and operated by a cooperation between the federal government and the federal states. PortalU regularly collects all environmental information of the state authorities in Baden-Württemberg available on the internet and provides the public with different options to arrange and search these data.

A main characteristic of PortalU is that the public data provided by the various authorities all over Germany are described in a uniform manner in the Environmental Data Catalogue. The data descriptions (metadata) from the UIS BW naturally can be found in this catalogue. At present it is not the intention of PortalU to provide data from the municipalities. This gap is filled by the Environmental Information Portal of Baden-Württemberg (Portal Umwelt-BW)[8],

[6] http://www.lubw.baden-wuerttemberg.de/servlet/is/58763/umweltdaten_2009_komplett.pdf?command=downloadContent&filename=umweltdaten_2009_komplett.pdf

[7] http://www.portalu.de/

[8] http://www.umwelt.baden-wuerttemberg.de/servlet/is/811/

in which all providers of environmental information on the state level as well as the larger cities in Baden-Württemberg take part with their specific websites. Stringent regulations prevent inconsistencies and duplication of work concerning PortalU. Therefore all information providers on the state level are held in the same way both in Portal Umwelt-BW and PortalU. The structuring, the metadata and the tagging of PortalU have been adopted by Portal Umwelt-BW.

Portal Umwelt-BW is a central entry point for thematically structured navigation and search facilities [9]which allows access to a large number information services. In the meantime this concept – including the administration tool and the search engine (Google Search Appliance, GSA[10]) – has been adopted by the responsible ministries of the Federal States of Saxony-Anhalt and Thuringia.

The GSA is a search engine with the functionality and performance of the well-known Google engine. This means, that search requests are composed by entering one or more search terms into the search form. The usage of extended search forms is also possible. Furthermore, extended search facilities can be configured, as illustrated in the following examples:

Search in databases: It is possible to access various database systems (e.g. Oracle, MySQL, DB2, Sybase) and to index directly the contents of database queries. URLs contained in database queries can be covered by the building of the search index. Web applications which are hidden behind complex selection lists or query forms can thus be found by means of full text search. Hits of web pages are obtained which do not contain the search terms, e.g. a search for "protection areas" in UDO also delivers maps of these protection areas. Figure 2.23 displays the source references provided in the context of the search for the term "Feinstaub" 'particulate matter'.

Key Matches: Key Matches are a possibility to assign relevant web pages directly to certain search terms. In the following example, a search for "respirable dust (Feinstaub)" offers a link "try here (versuchen Sie es einmal hier)", which leads to a page where measurement values for respirable dust are offered.

One Boxes: One boxes allow for the propagation of search queries to further systems. The results of these systems are expected within a configurable time limit in a generic format which covers various type of content. The results of a search can thus be enriched with further content.

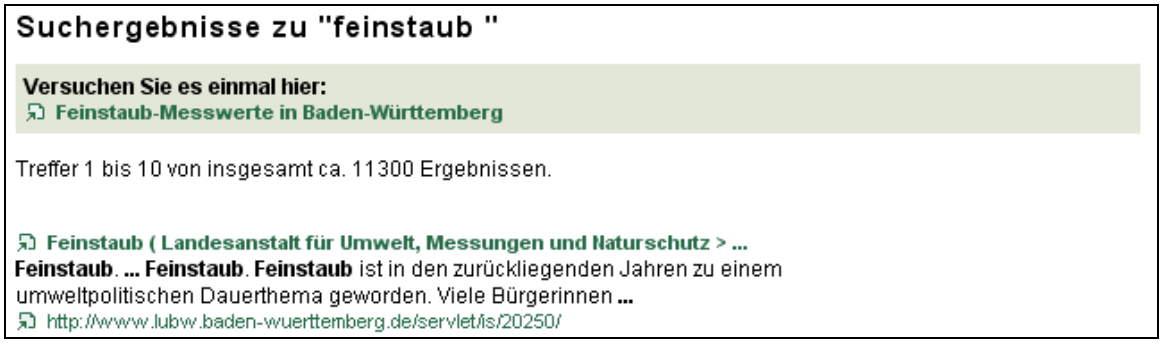


Figure 2.23: References provided in the context of the search for "Feinstaub" 'particulate matter'

2.2.3.4 Dissemination of Environmental Information through Web Services

[9] Thorsten Schlachter et al.: LUPO: Fortgeschrittene Suchfunktionen in den Landesumweltportalen von Baden-Württemberg, Sachsen-Anhalt und Thürigen. In: Mayer-Föll et. al. (ed.): F+E Vorhaben KEWA , Phase IV 2008/2009, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte FZKA 7500, S. 157-166

[10][10] http://www.google.de/enterprise/gsa/

Environmental information is disseminated by the use of web services at an increasing rate both to users within the administration and to public users over the internet. By means of the reporting system, any query can be transformed into a web service (according to the REST specification). Today about 60 standardised web map services (WMS) have been constructed (according to the standards of the Open Geospatial Consortium, OGC). Anyone running a web server, therefore, can integrate corresponding up-to-date maps from the UIS BW dealing with subjects such as nature conservation, water management and climate protection into their own website.

Due to the increasing amount of web services, provided by the UIS BW[11] to environmental offices in Baden-Württemberg, an automated registry had to be established. Consequently, a registry in conformity with the UDDI standard has been developed, especially for the reporting services. The web map services, which are at the same time developed for the Spatial Data Infrastructure of Baden-Württemberg (GDI-BW) according to the INSPIRE standards, are also registered in a special Catalogue of Metadata and Services (RIPS-OK), which conforms to ISO 19115.

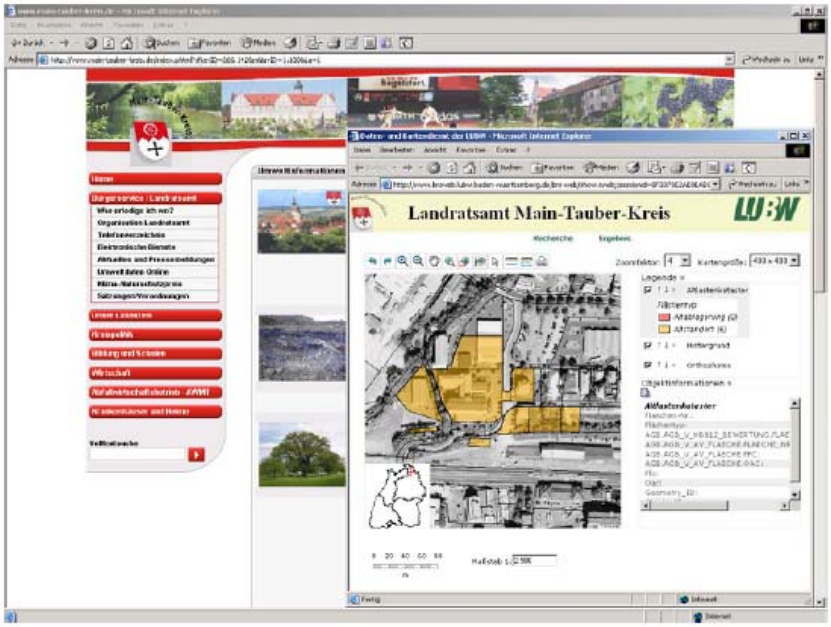


Figure 2.24: Integration of reports into local information provider websites

One service frequently used by environmental offices of the municipalities, concerns the reporting services of UDO. This service enables the environmental administrations of the district authorities to integrate reports generated using the corresponding data of the UIS reference database within the area of the district into the websites of the municipal information providers (as illustrated in Figure 2.24 above).

2.2.4 OTHER AIR QUALITY SERVICES

Globally, and especially within Europe, there are many web services which offer real-time or near-real-time air quality information. It is impossible to cover even a small part of them. Therefore, we restrict, in what follows, to a brief overview of two representative regional-level and one national-level services.

2.2.4.1 THE LONDON AIR QUALITY NETWORK

The London Air Quality Network *http://www.londonair.org.uk* (LAQN), managed by the Environmental Research Group at King's College London, gives on its front page (Figure 2.25) an at a glance view of the current air quality

[11] http://brsweb.lubw.baden-wuerttemberg.de/wiki/Hauptseite

situation in the London area according to the British air quality index (values from 1 to 10, four color classes: low, moderate, high and very high).

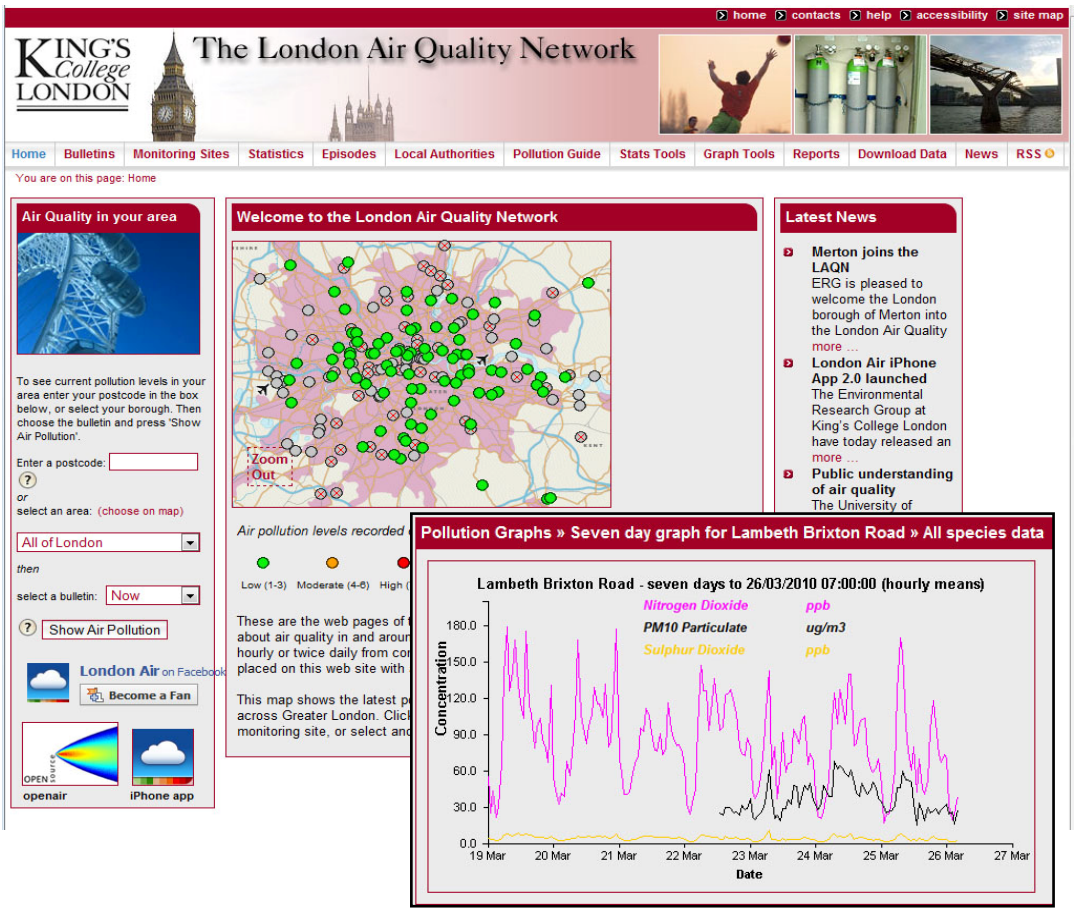


Figure 2.25: The front page of the London air quality service with the map presentation of current air quality according to an air quality index. An example of a graphical presentation of pollutant concentrations at one measuring site is given in the smaller overlay.

The data is updated hourly. The user can select different districts for closer inspection. The data of each measuring station can also be viewed in graphical form and downloaded (in csv format) for further use.

LAQN utilizes modern information technologies and also the social media. The air quality information can be viewed with an iPhone application (Figure 2.26) either on a map, by site or at selected sites. There are also push notifications to keep the user always up to date with the current situation. LAQN RSS provides feed of both news and of the latest air quality readings. The subscription can be tailored to inform about moderate or above readings at selected site(s) and local authorities. LAQN is also on Facebook and in Twitter it has two feeds (hourly update and daily summaries along with news items).

Figure 2.26: The iPhone application of LAQN keeps the user up to date of the air quality without the need of checking a website.

2.2.4.2 THE AIRPARIF SERVICE

AIRPARIF is the organisation responsible for monitoring air quality in the Paris region (Île-de-France). In addition to monitoring, the responsibilities of AIRPARIF include the forecasting of air pollution episodes, assessing the impact of emission reduction measures and informing the public authorities and general public about air quality. Their website [12]presents air quality characterization, according to the French air quality index (ATMO), for yesterday, today and tomorrow with illustrative graphics form (Figure 2.27). The ATMO index is assigned values from 1 to 10, but it is calculated differently from the above mentioned British index. The values are assigned in six classes (*très bon*, *bon*, *moyen*, *médiocre*, *mauvais*, *très mauvais*) and illustrated with either three colours or a sliding colour scale. Further information can be obtained by clicking either the index numbers or the thumbnail maps of the respective days. The observations at each station can be viewed as graphs or tables and downloaded (in csv format) for further use (Figure 2.28).

[12] http://www.airparif.asso.fr

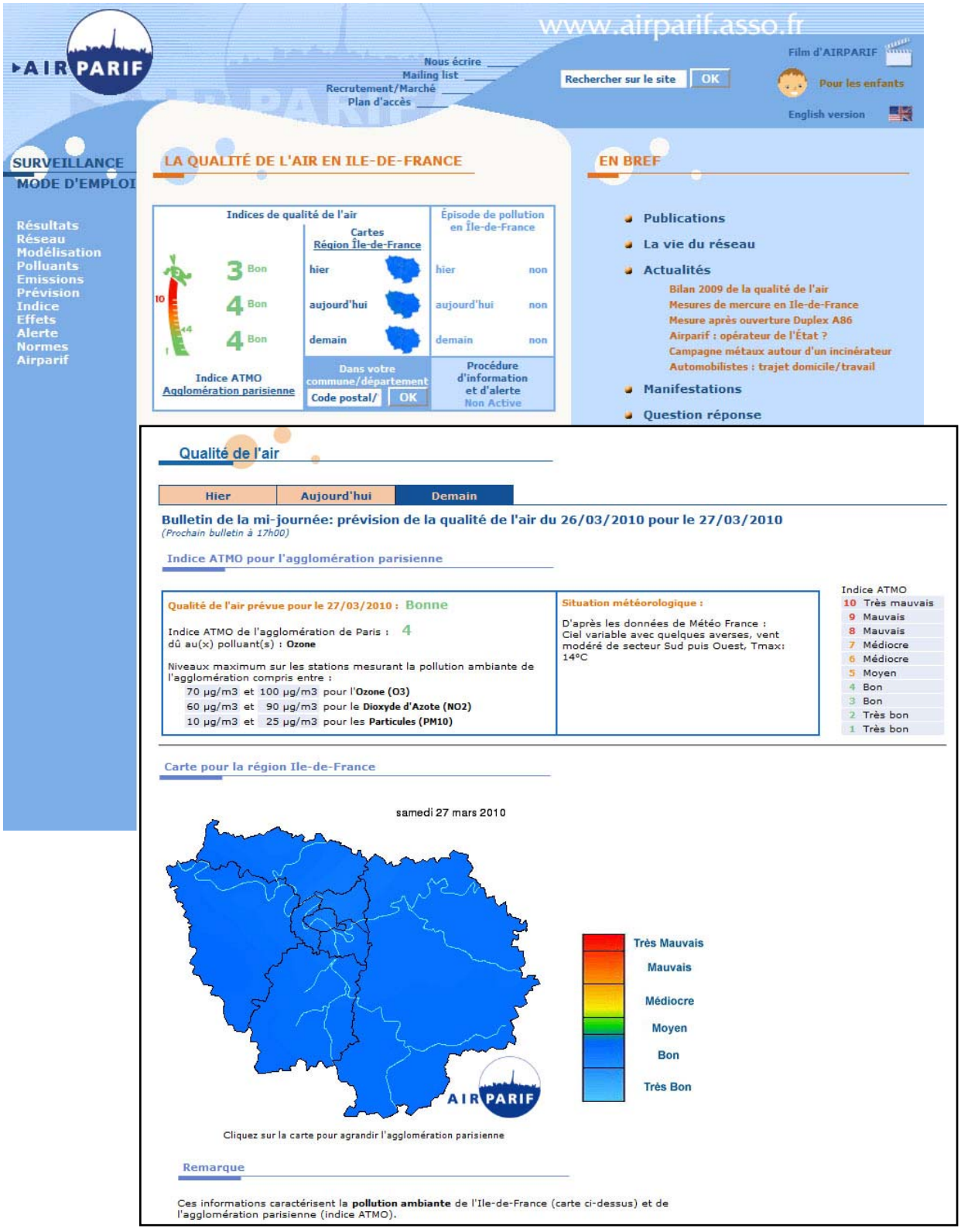
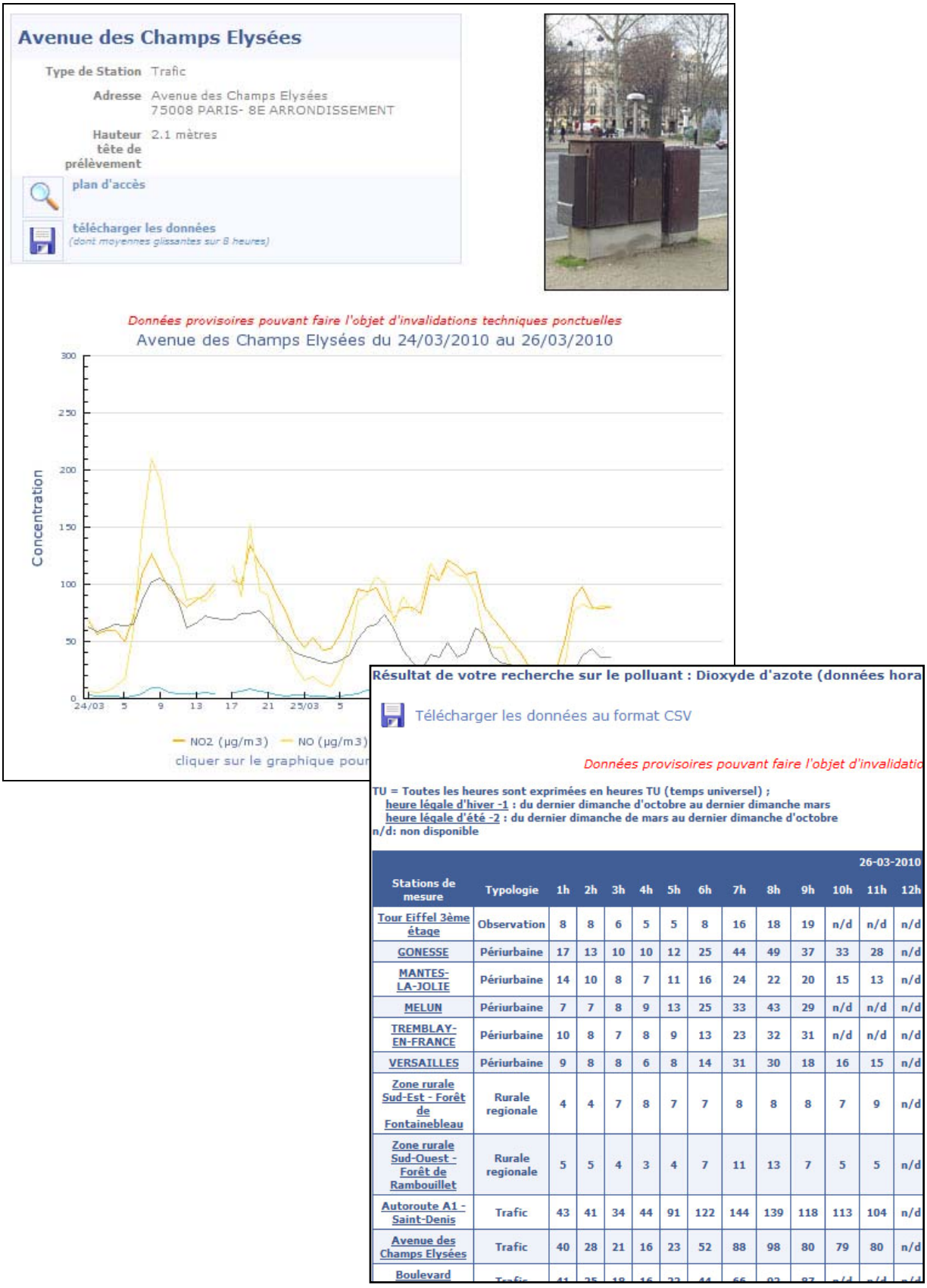


Figure 2.27: The front page of the AIRPARIF web service with an overlay of the forecasted air quality for the next day.



26-03-2010

Stations de mesure	Typologie	1h	2h	3h	4h	5h	6h	7h	8h	9h	10h	11h	12h
Tour Eiffel 3ème étage	Observation	8	8	6	5	5	8	16	18	19	n/d	n/d	n/d
GONESSE	Périurbaine	17	13	10	10	12	25	44	49	37	33	28	n/d
MANTES-LA-JOLIE	Périurbaine	14	10	8	7	11	16	24	22	20	15	13	n/d
MELUN	Périurbaine	7	7	8	9	13	25	33	43	29	n/d	n/d	n/d
TREMBLAY-EN-FRANCE	Périurbaine	10	8	7	8	9	13	23	32	31	n/d	n/d	n/d
VERSAILLES	Périurbaine	9	8	8	6	8	14	31	30	18	16	15	n/d
Zone rurale Sud-Est - Forêt de Fontainebleau	Rurale regionale	4	4	7	8	7	7	8	8	8	7	9	n/d
Zone rurale Sud-Ouest - Forêt de Rambouillet	Rurale regionale	5	5	4	3	4	7	11	13	7	5	5	n/d
Autoroute A1 - Saint-Denis	Trafic	43	41	34	44	91	122	144	139	118	113	104	n/d
Avenue des Champs Elysées	Trafic	40	28	21	16	23	52	88	98	80	79	80	n/d
Boulevard	Trafic	41	25	18	16	22	44	66	92	87	n/d	n/d	n/d

Figure 2.28: The viewing and downloading air quality observations (also provisional data) from the AIRPARIF service is made very easy.

2.2.4.3 THE NILU SERVICE

The Norwegian Institute for Air Research (NILU) maintains a very straightforward and uncomplicated web service for air quality in Norway at *http://www.luftkvalitet.info*. The front page (Figure 2.29) presents an overview of the current air quality according to the Norwegian air quality index. The index has four classes (very good, good, bad, very bad) and four colour codes. Again, the calculation of the index differs from both the British and the French indices. The data is updated hourly.

Figure 2.29: The front page of the NILU air quality service presents the air quality according to the Norwegian air quality index. Different types of measuring stations are distinguished by different symbols (filled circle: local station, open circle: regional station).

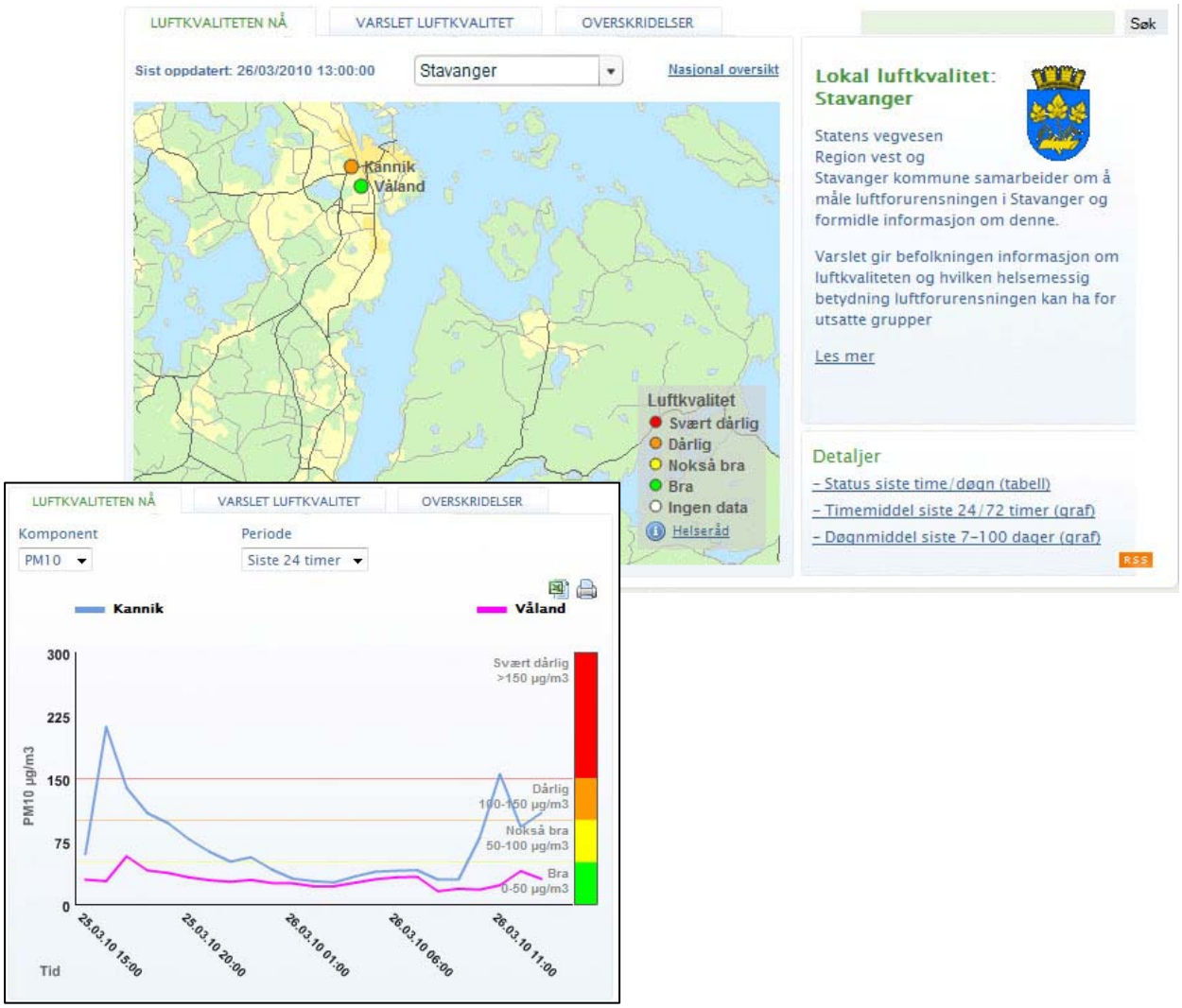


Figure 2.30: The observations at each monitoring station can be viewed as graphs with variable time scales.

The visitor can zoom to the region or place of interest by clicking on the map. The observations at each monitoring station can be viewed as tables or graphs and downloaded as Microsoft Excel files for further use (Figure 2.30). Download of historical (calibrated or validated) data is available from the service for registered users.

In addition to real time data the NILU service offers air quality forecasts for the current day and the next day. The service is available during the winter period The forecast is illustrated either by a table with the air quality index colour for three times a day (hours 10, 14, and 18) or with an hourly graph over the forecast period (Figure 2.31).

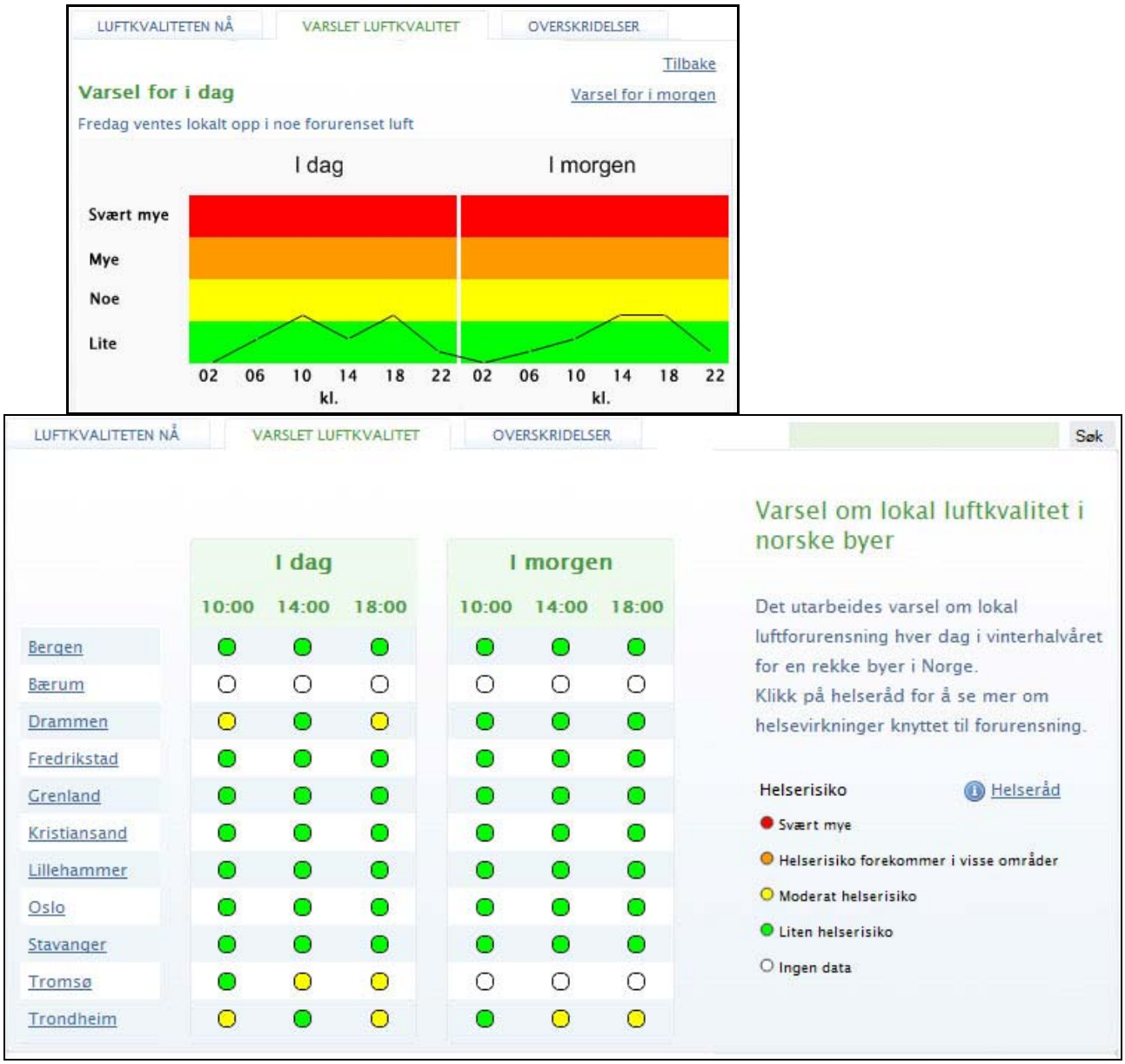


Figure 2.31: Air quality forecasts are presented for the current day and the next day using the air quality index.

2.3 Prototypical Services as Outcome of R&D Projects

A number of innovative prototypical environmental services have been presented in the recent past – some of them in the development state, others advanced and (nearly) operational, and still others run in the experimental mode for a while and then stopped or reduced to the provision of basic services. Compared with the operational services, the innovations lie first of all in the following features:

(i) use of cutting edge forecast models; cf., e.g., the Helsinki TestBed above; [Lohmeyer et al., 2007; Bronder et al., 2007 and others];

(ii) delivery of information via a variety of different communication channels – in particular, mobile services, and email; cf., e.g., [Rose et al., 2004].

(iii) tailoring the amount and kind of the information to the profile of the user; cf., e.g., [Karatzas et al., 2006; Bouayad-Agha et al., 2006, 2007].

(iv) use of state-of-the-art information production techniques; cf., e.g., [Busemann and Horacek, 1997; Coch, 1998; Wanner et al., 2007];

(v) adapting a cross-border view on environmental conditions and their presentation; cf., e.g., MARQUIS [Wanner et al., 2007].

The reasons why many innovative prototypical implementations did not fid their way into a stable operational mode are manifold and certainly differ from case to case. A thorough analysis is needed in the context of PESCaDO in order to be able to take precautionary measures from the beginning and avoid this situation.

In what follows, we focus on two prominent prototypical services: MARQUIS (EDC-11258) and APNEE / APNEE-TU (IST-34154). MARQUIS and APNEE have been chosen for presentation because they present, to the best of our knowledge, the most advanced models for environmental information communication to the citizen. To be monitored in the future are also ongoing initiatives such as GENESIS (FP7-IST-223996, http://genesis-fp7.eu/), which addresses the problem of environmental information management from the perspective of health impact. Since no concrete information is available as yet on the technologies in GENESIS, we sketch it below merely along generic lines.

2.3.1 MARQUIS

MARQUIS (Multimodal Air Quality Information Service for General Public), funded by the EC in the framework of the eContent Programme (EDC-11258) targeted the development of a highly multilingual cross-border air quality (AQ) information service for five of European regions: Baden-Württemberg, Catalonia, Finland, Portugal and Upper Silesia. The languages of the service were: Catalan, English, Finnish, French, German, Polish, Portuguese, and Spanish. Consider, for illustration, fragments of an English report generated by MARQUIS:

> *The air quality index is 3, which means that the air quality is satisfactory. This is due to the ozone concentration. The NO2 concentration, the SO2 concentration and the PM10 concentration do not have any influence on the air quality. The current air quality index (3) is today's highest: the lowest air quality index was 2 (at midnight). Between midnight and 7AM, the air quality index remained stable at 2 and between 8 AM and 9PM, it remained stable at 3. … The PM10 is low. This is due to rainy weather conditions. The nitrogen dioxide concentration (3μg) is also very low. Thus, no harmful effects to human health are expected. The first NO2 concentration measured today (at 1AM) was still 29μg. … The sulfur dioxide concentration remained stable at 1μg between 8AM and noon and between 1PM and 9PM it remained stable at 0μg.*

The most innovative features of MARQUIS were: (i) coverage of all pollutant substances measured in the regions in question; (ii) the use of cutting-edge AQ forecast models and of AQ assessment and interpretation models; (iii) user-orientation: a fine-grained user profile typology has been developed and made accessible by individual users for personalization, such that the air quality bulletins could be tailored to the specific needs of the users; (iv) use of advanced computational linguistic technologies for the generation of the bulletins; (v) coverage of a wide range of communication channels: email, web, mobile phone services (SMS, MMS), press, TV.

After completion of the project, the service has been further developed to serve better for the generation of air quality bulletins in Finland. It is still used in a reduced form in Upper Silesia. In the other involved regions, MARQUIS has not been put to operational service due to uncertain market prospects and lack of further funding.

2.3.1.1 Architecture of the MARQUIS-Service

The diagram in Figure 2.32 shows the architecture of MARQUIS, which is a “two pipe” architecture. The first pipe consists of the data processing pipe; the second pipe is the information request and delivery pipe.

The date processing pipe has the following functions:

1. Monitoring of air pollutant concentrations and of meteorological conditions in the five MARQUIS-regions and execution of data quality assurance and air pollution forecast models.
2. Delivery of the measured and forecasted data from the local DBs to the MARQUIS-server.
3. Assessment and interpretation of the delivered data with respect to their relevance to any of the EU and regional environmental legislation issues and to any of the MARQUIS-users; determining the primary meteorological and contextual influence on the measured and forecasted air quality (for explaining / justifying them). Air quality forecast models may also be run in this stage when required.

The information request and delivery pipe looks as follows:

1. Receiving an information request from a user via the MARQUIS-Client interface (this can be an automated periodic request, a request generated upon the exceedance of a threshold (either personal or legal) by the concentration / index of a specific pollutant or by the air quality index, or a request launched by the user).
2. Selecting the content that is relevant for the user in question from the structure produced in step 3 of pipe 1.
3. Generating the discourse structure of the content to be conveyed to the user, determining the appropriate mode for the individual chunks of the content, and starting the corresponding information generators.
4. Generating the information by the table, graphic and multilingual text generator.
5. Conveying the generated information to the user using his/her preferred communication channel.

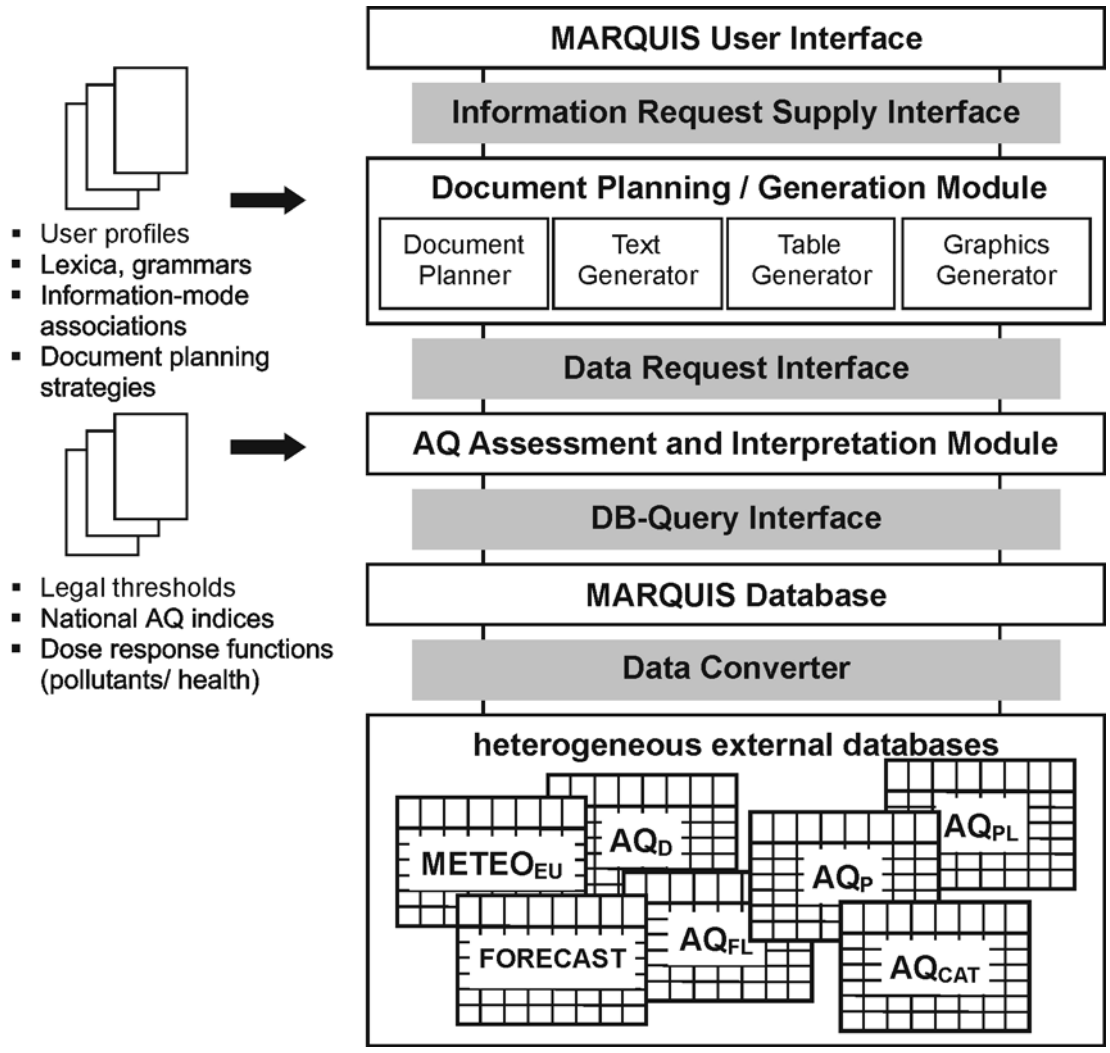


Figure 2.32: Architecture of the MARQUIS service

Unlike any other environmental service we are aware of, MARQUIS contains a full-fledged multimodal document generator.

2.3.1.2 USER PROFILE TYPOLOGY IN MARQUIS

One of the most distinctive features of MARQUIS is its fine-grained default user profile typology, which can be further personalized with respect to the type and amount of desired information and the mode and channel of its delivery. For the definition of the default user profile typology in the context of AQ information, MARQUIS draws upon three dimensions; cf. also [Molina et al., 2005]: (i) expertise with respect to air pollution, (ii) air pollution sensitivity of the target audience, and (iii) the preferred communication channel. The expertise dimension covers the specification of the kind of background information needed by a user, to what extent the AQ information should be qualitative (and thus easier to understand) or quantitative (and thus require further interpretation), and in which mode the information is preferrably to be presented (text, table, or graphic)—provided the communication channel chosen by the user allows for variation of the mode. The air pollution sensitivity dimension links the information delivery to specific AQ levels and concentrations of different pollutant substances, include or omit health warnings, mention or not the weather conditions, etc. The communication channel dimension determines further the mode of presentation, the conciseness (i.e., the amount) of the information offered, etc. The default user profile typology reflects these three dimensions:

1. domain professional
2. medical professional
   - 2.1 respiratory disease specialist
   - 2.2 heart specialist
   - 2.3 general medical professional
3. public
   - 3.1 general public
   - 3.2 outdoor active public
   - 3.3 patient
     - 3.3.1 respiratory disease patient
     - 3.3.2 heart patient

Obviously, this typology can be further extended – for instance, children are likely to form another category which is to be taken into account. After registration, the user can *de facto* change all parameter settings – except a few that are basic for each profile. Figure 2.33 shows some of the default settings for the profile "general public" with the web as the communication channel.

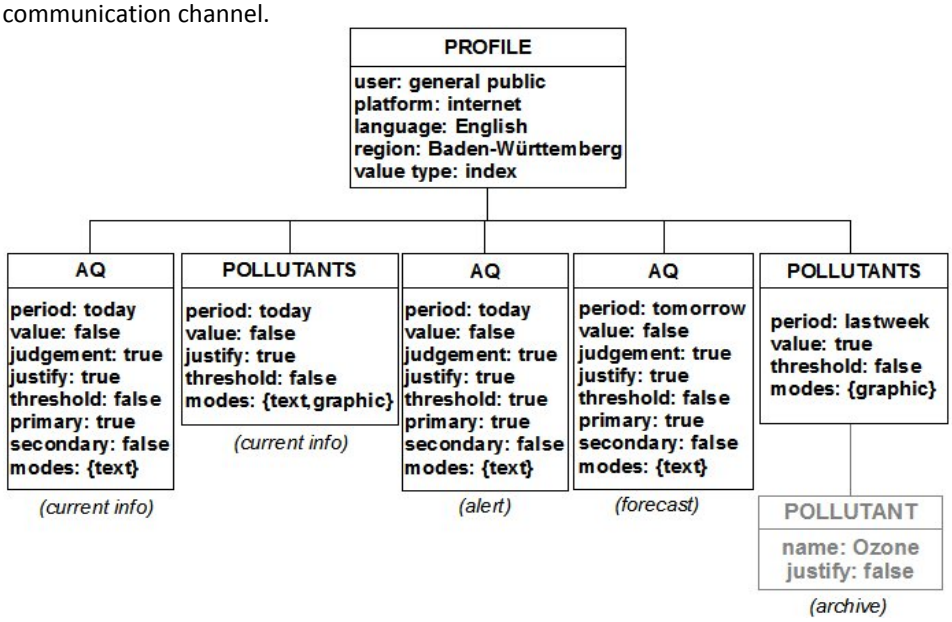


Figure 2.33: Default settings for the general public profile, internet (simplified)

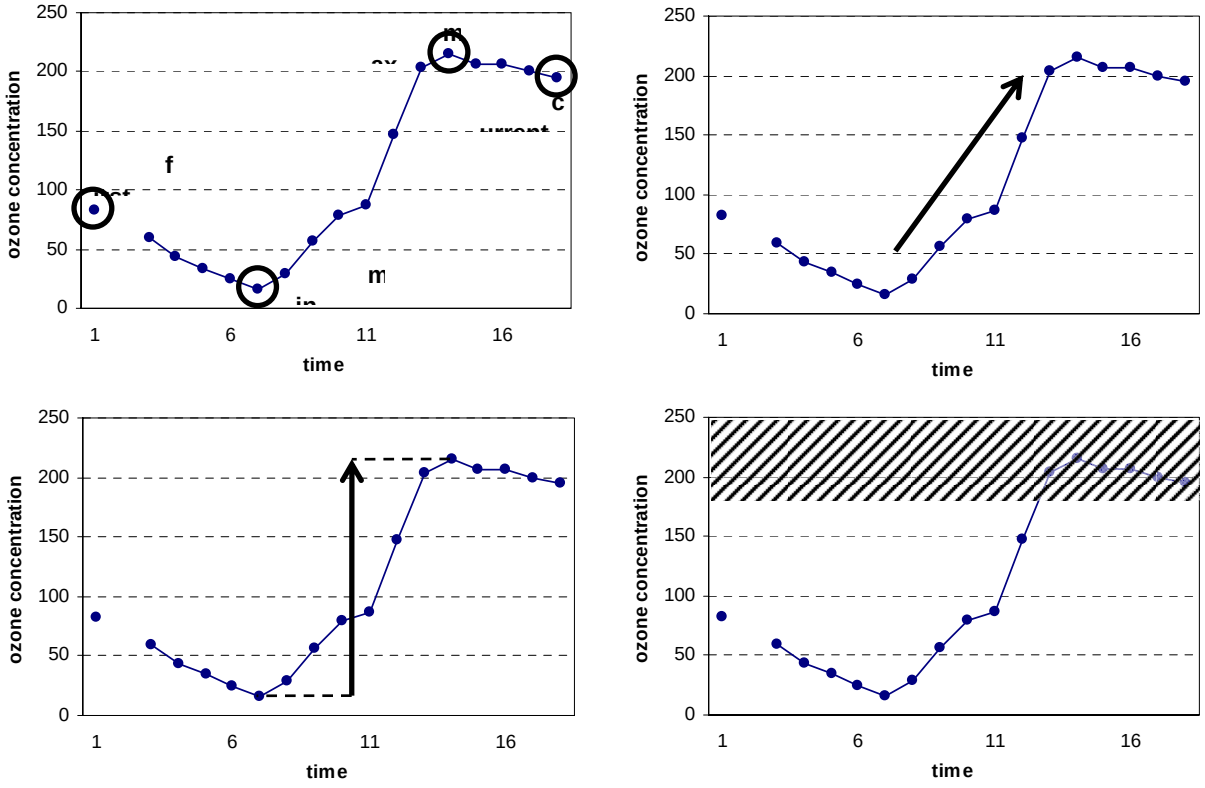


Figure 2.34: Analysis of the pollutant curves in MARQUIS

2.3.1.3 AQ Assessment and Interpretation in MARQUIS

Another innovative feature in MARQUIS is the assessment and interpretation of the raw measured and forecasted air quality data. The assessment ranges from the calculation of the air quality / pollutant indices and their association with region-specific qualitative value scales (such, as, e.g., excellent, good, satisfactory, bad, very bad), over (legally predefined) health impact assessment to a detailed analysis of the pollutant curves. For the calculation of the indices and for valuing of the measured / forecasted pollution, the scales of the region of origin of the user are taken into account – such that the user receives information in the same scales as he/she is used to from their own region. This feature makes MARQUIS a "cross-border" service.

The analysis of the pollutant curves retrieves information concerning the maxima and minima, abrupt changes, threshold exceedance, and the like. Consider Figure 2.34 above for illustration and consult [Nicklass et al., 2007] for details on the way the assessment and interpretation module in MARQUIS works.

2.3.1.4 Bulletin Generation in MARQUIS

As mentioned above, the generation of air quality in MARQUIS involves a state-of-the-art document generator, which starts from the "assessment plan" as produced by the assessment and interpretation module.

Particularly interesting for PESCaDO is in this context the multilingual text generator, which is divided into two submodules: the text planning submodule and the linguistic generation submodule. Figures 2.35 and 2.36 illustrate the way both of them work. Cf. [Bouayad-Agha et al., 2006, 2007] for the description of the text planner and [Bohnet et al., 2007] for the description of the linguistic generator. See also Section 11 for more details on the delivery of textual information.

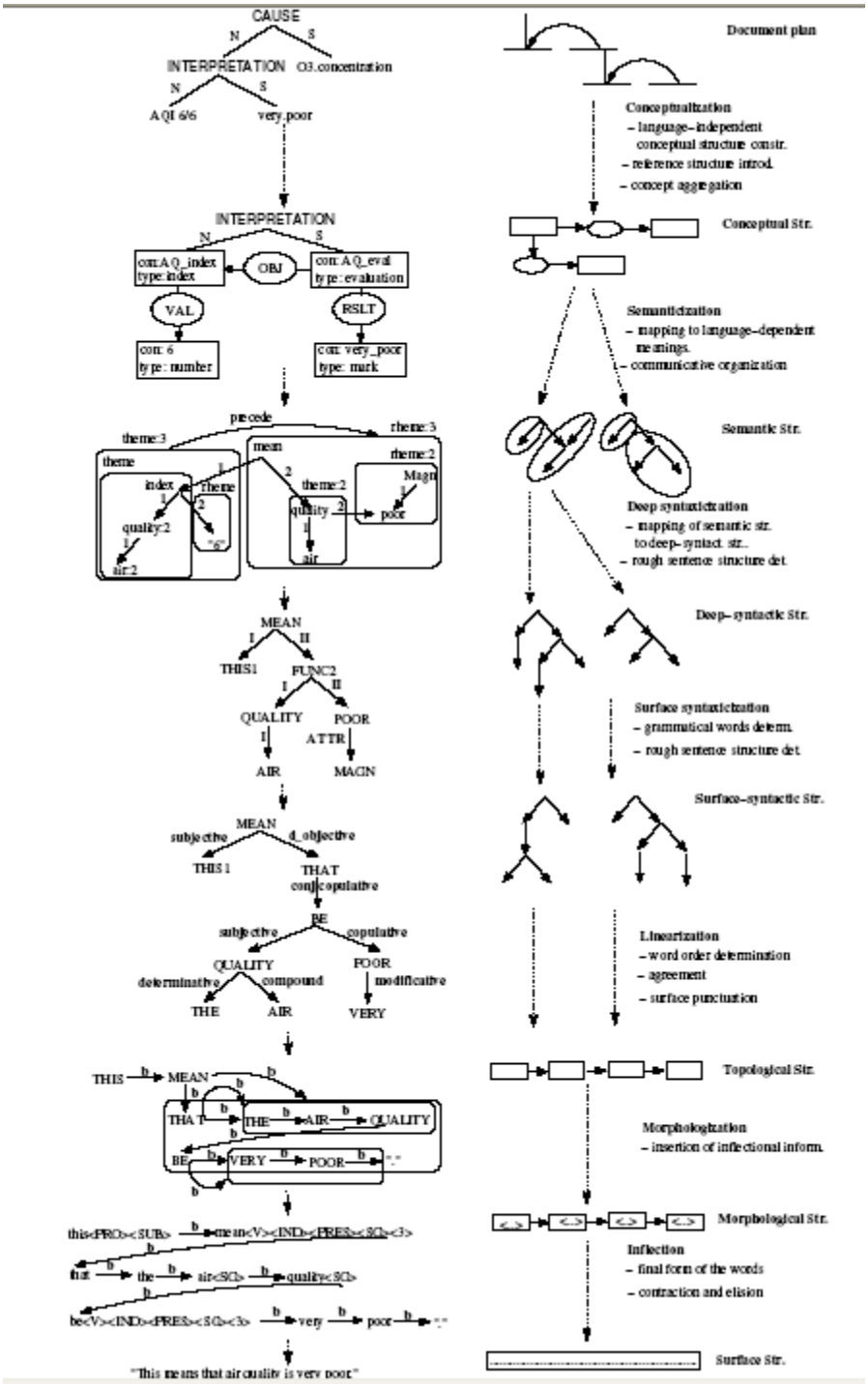


Figure 2.35: Architecture of the linguistic generation submodule in MARQUIS

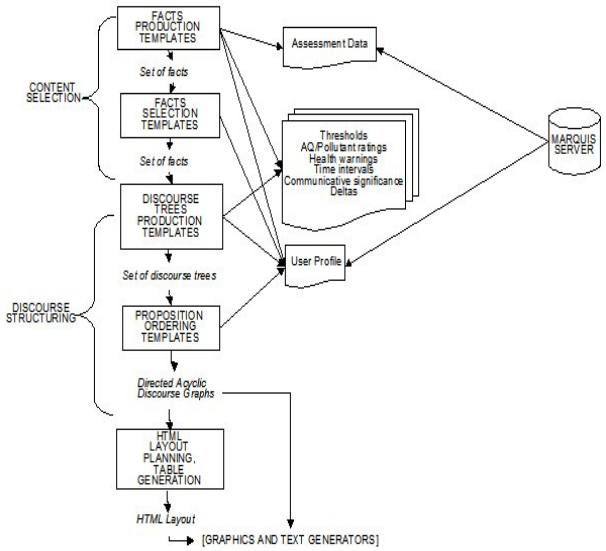


Figure 2.36: Architecture of the text planning submodule in MARQUIS

2.3.2 APNEE

APNEE and its successor APNEE-TU (IST-34154) were two projects funded by the EC in FP5 from 1999 to 2004. The focus of this pair of projects was on the coverage of a wide range of geographical regions in Europe and delivery of information via a range of different communication channels. Consider the dissemination platform of APNEE and the communication channels served by APNEE (from [Rose et al., 2002]) in Figure 2.37.

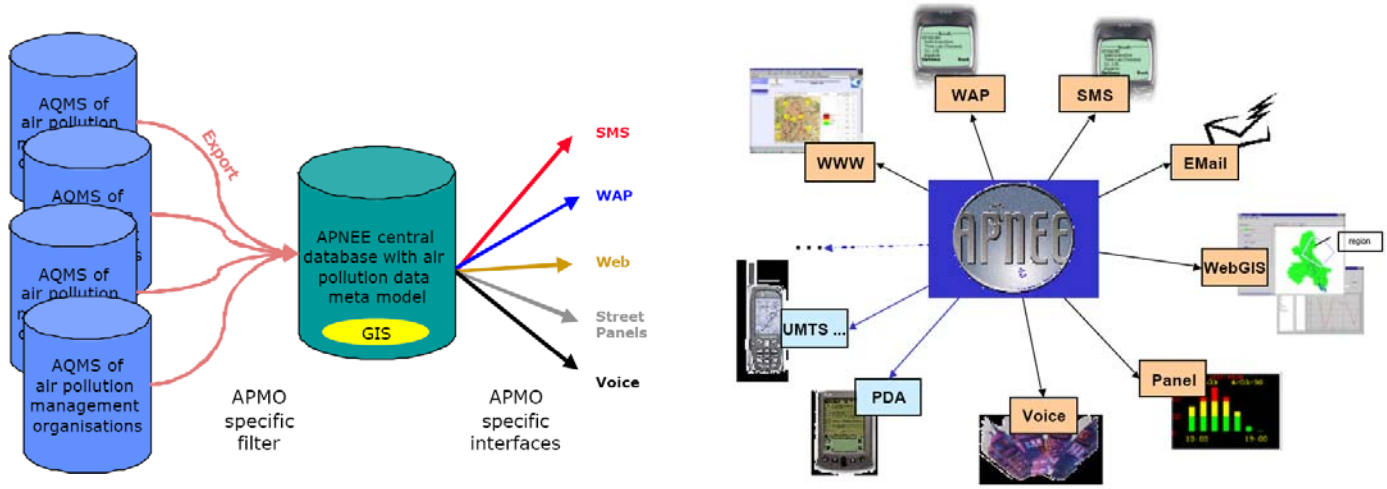


Figure 2.37: APNEE's dissemination platform and communication channels [Rose et al., 2002]

To be highlighted is the innovative push service in APNEE that tracks the user's location via the mobile phone connection and provides information on air quality in this location. While some citizens may object to being monitored via mobile phone connections due to privacy considerations and concerns of interception, a similar GPS-based service may be also a good option for PESCaDO.

Although nearly all APNEE(-TU) publications [Rose et al., 2002; Peinel et al., 2003; Peinel and Rose, 2004; Karatzas et al., 2006] state that the information provided by the APNEE-service is user-tailored, it seems that, unlike in MARQUIS, user-orientation is understood uniquely in the sense that the user can subscribe and select the information that he/she desires to receive as well as the communication channel.

Also in contrast to MARQUIS, which focuses on textual information, considering graphics and tables as auxiliary modes, APNEE argues for communication of environmental information via pictograms (Figure 2.38) colored map and colored number displays (Figure 2.39). The similarity to the NILU service (2.2.4.3 above) is not by chance: NILU participated in APNEE as one of the leading partners of the consortium.

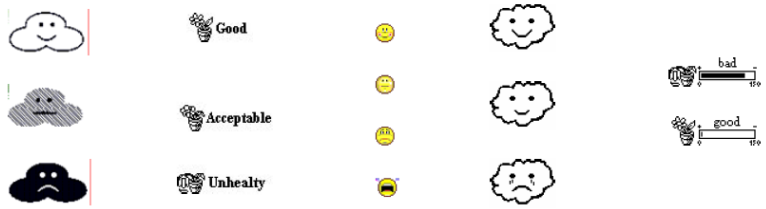


Figure 2.36: Use of pictograms in APNEE for communication of air quality information

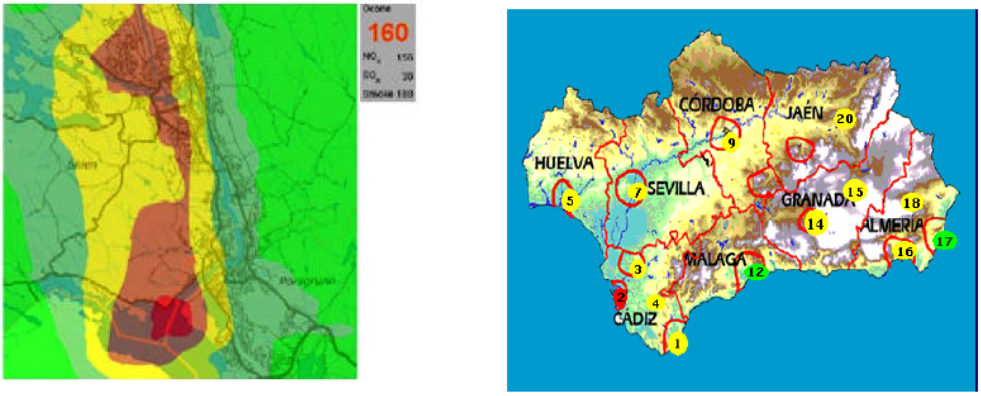


Figure 2.39: Use of colors in APNEE for displaying air quality information

The APNEE-Consortium performed trials in Norway (Grenland and Oslo), Germany (Stuttgart and Dresden), France (Marseille), Spain (Madrid, Andalusia, and Canary Islands), and Greece (Thessaloniki and Athens). To the best of our knowledge, this has been the biggest scope of a user-oriented environmental information delivery project so far.

2.3.3 GENESIS

GENESIS (IST-223996) envisages three major pilots related to air quality information management, fresh water quality management and coastal water quality management. The schematic outline of the pilot services as sketched in [GENESIS-D3100.3, 2009] is shown in Figure 2.40. The diagram suggests that the focus of GENESIS is indeed on the integration of different data sources and the processing of data in accordance with existing air quality and health regulations – and not on the provision of user-tailored personalized information that aims to support the user in his/her decision making. Furthermore, it seems obvious that GENESIS does not aim to use any advanced reasoning or inference technologies.

The specification of the pilots in [GENESIS-D8100.3, 2009; GENESIS-D3100.3, 2009] furthermore reveals a prominent role of the user not only as consumer, but also as provider of information (e.g., data of the occurrences of jelly fish in a specific region) that is incorporated into the system DB and digested (for instance, for the production of risk index maps in the case of jelly fish occurrences). This is in accordance with the theme of the project, namely environmental information management.

No further detailed information is yet available on GENESIS. A continuous monitoring of its website and contacts to its consortium will be essential for PESCaDO in case GENESIS (as an IP) decides to incorporate, for instance, decision support techniques.

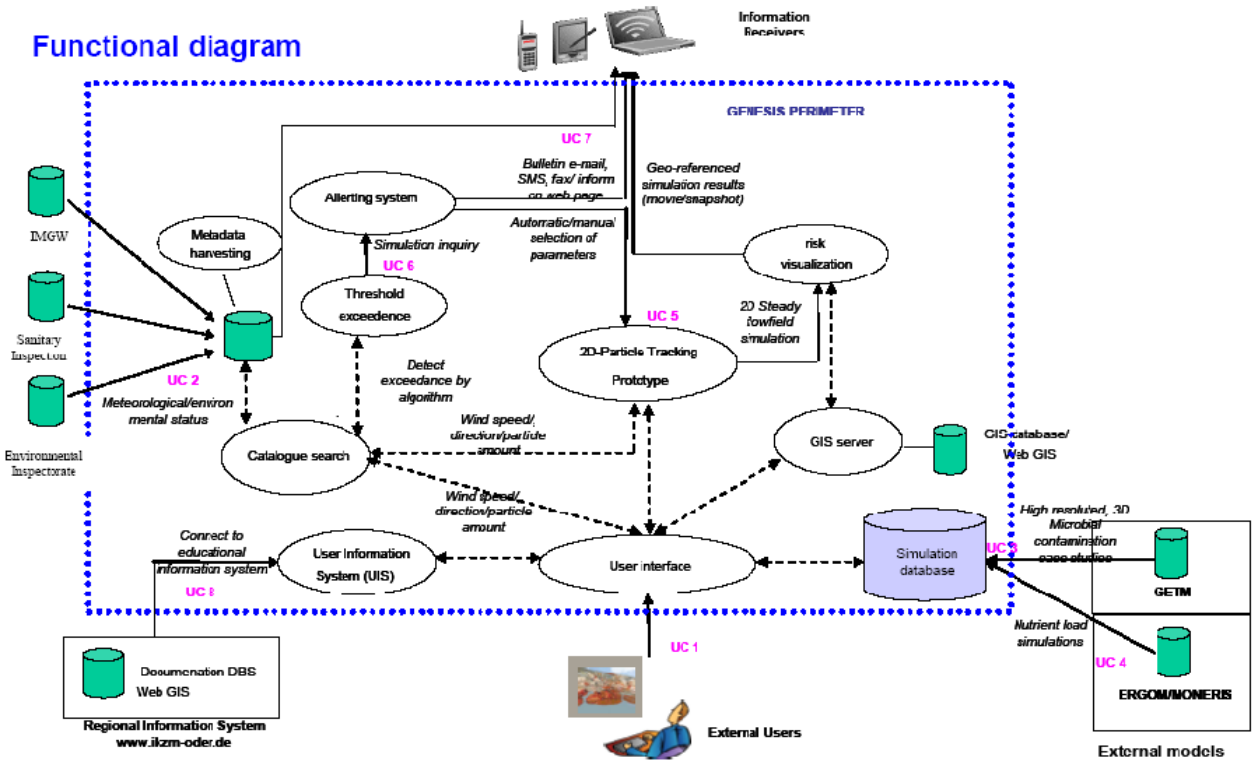


Figure 3.40: Diagram of pilot use cases in GENESIS

2.4 REFERENCES

[Bohnet et al., 2007] B. Bohnet, F. Lareau and L. Wanner. `Automatic Production of Multilingual Environmental Information'. In Proceedings of EnviroInfo 2007, Warsaw, 2007.

[Bouayad-Agha et al., 2006] N. Bouayad-Agha, L. Wanner, and D. Nicklass. Discourse Structuring of Dynamic Content. In Proceedings of the Spanish Conference on Computational Linguistics (SEPLN), Zaragoza, 2006.

[Bouayad-Agha et al., 2007] N. Bouayad-Agha and L. Wanner. `Text Planning of Air Quality Information'. In Proceedings of EnviroInfo 2007, Warsaw, 2007.

[Bronder et al., 2007] Bronder, J., C. Klis and J. D lugosz. 2007. Air quality information service in Upper Silesia. Proceedings of the EnviroInfo Conference. 31–39. Warsaw.

[Busemann and Horacek, 1997] S. Busemann and H. Horacek. 1997. Generating air-quality reports from environmental data. Proceedings of the DFKI Workshop on Natural Language Generation. 15–21. Saarbrücken, Germany.

[Coch, 1998] Coch, J. 1998. Interactive generation and knowledge administration in MultiMeteo. Nint International Workshop on Natural Language Generation. 300–303. Niagara-on-the-Lake, Ontario, Canada.

[GENESIS-D8100.3, 2009] GENESIS, ACR Pilot Specifications http://genesis-fp7.eu/publications

[GENESIS-D3100.3, 2009] GENESIS, SGH Pilot Specification http://genesis-fp7.eu/publications

[Karatzas et al., 2006] K.Karatzas, G. Endregard, and I. Floisand. 2006. Citizen-oriented environmental information services: Usage and Impact. In EnviroInfo 2005, Environmental Communication in the Information Society, 19th International Conference "Informatics for Environmental Protection", Networking Environmental Information, September 7 - 9, 2005, Masaryk University in Brno, Czech Republic.

[Keitel et al., 2009] Adree Keitel, Roland Mayer-Föll, Albrecht Schultze: Framework Conception for the Environmental Information System of Baden-Württemberg. Proceedings of the European Conference TOWARDS eENVIRONMENT, March 25-27, 2009, pages 461-468. Prague, Czech Republic.

[Lohmeyer et al., 2007 A. Lohmeyer, I. Düring, M. Giereth, T. Hoffmann, D. Nicklass, M. Klein, H. Scheu-Hachtel, C. Sörgel and L. Wanner. `Kurzfrist-Feinstaub-Immissionsprognose mit den Systemen MARQUIS und ProFet' Gefahrstoffe. Reinhaltung der Luft, 67(7/8): 319-326, 2007.

[Molina et al., 2005] Molina, T., A. Panighi and A. Flores. 2005. MARQUIS (EDC-11258), Deliverable 3.1: Specification of the User Profiles. Technical report: TVC Netmedia.

[Nicklass et al., 2007] D. Nicklass, N. Bouayad and L. Wanner. `Addressee-Tailored Interpretation of Air Quality Data'. In Proceedings of EnviroInfo 2007, Warsaw, 2007.

[Peinel et al., 2003] G. Peinel, T. Rose, M. Sedlmeyr et al. 2003. APNEE – Air Pollution Network for Early Warning and Information Exchange in Europe. In Proceedings of the Third International Symposium Digital Earth, Brno.

[Peinel and Rose, 2004] G. Peinel and T. Rose. Dissemination of Air Quality Information: Lessons Learned in European Field Trials. In Proceedings of EnviroInfo 2004, Geneva, 2004.

[Rose et al., 2002] T. Rose, G. Peinel, M. Sedlmayr and K. Karatzas. "Citizen-centred Dissemination of Air Quality Information on Multi-modal Information Channels: the APNEE Project". In Euro-Sustain 2002, Rhodes, Greece. 2002.

[Schlachter et al., 2009] T. Schlachter et al.: LUPO: Fortgeschrittene Suchfunktionen in den Landesumweltportalen von Baden-Württemberg, Sachsen-Anhalt und Thüringen. In: Mayer-Föll et. al. (ed.): F+E Vorhaben KEWA , Phase IV 2008/2009, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte FZKA 7500, S. 157-166

[Wanner et al., 2007] L. Wanner et al. `From Measurement Data to Environmental Information - MARQUIS - A Multimodal AiR QUality Information Service for the General Public'. In Proceedings of Environmental Software Systems: Dimensions of Environmental Informatics, Vol. 7, Prague, 2007.

3 DISCOVERY OF ENVIRONMENTAL SERVICE NODES

3.1 GLOSSARY – ABBREVIATIONS

AJAX	Asynchronous JavaScript And XML
API	Application Programming Interface
ANN	Artificial Neural Network
BPEL4WS	Business Process Execution Language for Web Services
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
REST	Representational State Transfer
SOAP	Simple Object Access Protocol
SOM	Self-Organizing Map
TCP/IP	Transmission Control Program/Internet Protocol
UDDI	Universal Description Discovery and Integration
URL	Uniform Resource Locator
WS	Web services
WSDL	Web Service Description Language
XML	EXtensible Markup Language
XML-RPC	EXtensible Markup Language - Remote Procedure Call

3.2 DESCRIPTION OF THE PROBLEM DOMAIN

The problem of the discovery of environmental nodes in the Web can be considered as a very interesting challenge, as environmental information is highly distributed and available in heterogeneous forms. As environmental service nodes we consider any web page or web service that can provide environmental information. The discovery of environmental nodes in the web could be considered as a problem of a domain-specific web search, related also to web service discovery and to grid distributed systems.

The research on domain-specific search engines is a well-established area. Several techniques have been applied to develop such engines based on web searching, web crawling and query expansion techniques. More specifically, in the proposed methodologies either existing search engines are employed to access the web or predefined web sites are set as starting points. In the first case, queries are formulated and expanded by introducing terms that describe the domain. On the other hand, when a set of predefined web sites is available, different kind of crawlers (i.e. based on machine learning, ontologies, etc.) are applied depending on the exact methodology. In addition, post analysis of the retrieval results and filtering models can be applied to improve the precision. To the best of our knowledge, no environment-domain search engines are available so far.

Nowadays some of the environmental nodes exist in the form of web services. At the early years of service-oriented computing, web service discovery was mainly performed by scanning services registries. Although existing centralized registries could provide effective results, they have problems associated with centralized systems, so other approaches focused on having multiple registries grouped into registry federations. In addition, research is also conducted towards searching Web services on the Web either by employing regular search engines such as Google, Yahoo, or by developing specific crawlers that target service registries.

Finally, the problem of the discovery of environmental service nodes in the Web can be considered also common to modern open distributed computer systems, which offer various types of services [27]. However, most of the grid discovery approaches presuppose that the nodes to be searched post their functional “fingerprints”, i.e., their functional capacity and coverage, for external inspection. In the case of distributed environmental service nodes in the web – for instance, competing weather service nodes or complementary Air Quality (AQ) service nodes – this cannot be assumed.

3.3 DOMAIN SPECIFIC SEARCH ENGINES

Currently, generic search engines are the most popular access points for users on the Internet; however, the richness of the web content has made it progressively more difficult for users to acquire the desired information, especially when searching in domains that they lack domain-specific search knowledge. Thus, the development of a specialized search engine constitutes an imperative need. Based on the literature we will present the main methodologies for the implementation of a domain specific search engine:

Utilizing existing search engines [5]

Crawling the web [5]

In the next sections, the aforementioned techniques will be described in detail.

3.3.1 UTILIZING EXISTING SEARCH ENGINES

This methodology utilizes general purpose search engines to submit queries that are enriched with terms that characterize the desirable domain. Instead of employing a typical search engine, it is also possible to use a meta search engine, which is capable of merging the results of different search systems. The two basic approaches that employ existing search engines are illustrated in Figures 3.1 and 3.2. In the first case (Figure 3.1), the query is generated by applying machine learning techniques in order to extract terms (i.e. keyword spices) from positive and

negative sample web pages. In the second schema (Figure 3.2), the results obtained from the general purpose search engines by submitting queries consisting of terms that describe the domain (empirically selected), are filtered with filtering models and post-analysis of the retrieval information.

3.3.1.1 SEARCH ENGINES

The majority of existing generic web search engines possesses full-text searching capabilities. According to a report released in 2009 [8], the most popular search engines are Google Search, Yahoo! Search and Bing Search. The aforementioned search engines have developed APIs that provide external access to data and functionality of the services provided by them. Such APIs can be employed in the framework of Figures 3.1 and 3.2 to support automatic generation of results for new queries.

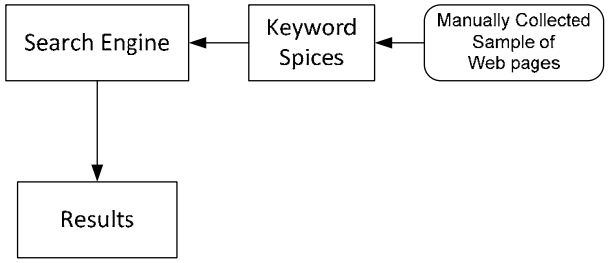


Figure 3.1: Keyword Spices

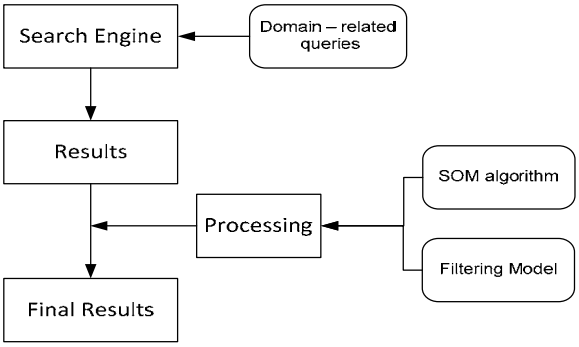


Figure 3.2: Analysis of search engines results

More specifically, Google has created the Google AJAX Search API, in order to allow users and developers perform inline searches over a number of Google services. This API was implemented both as a JavaScript library and as a RESTful interface that returns JSON encoded results. However, the use of Google AJAX Search API is restricted due to the limitations imposed by Google regarding its terms of use [9] (i.e. inability of modifying and storing the produced results, etc.). Yahoo! released BOSS (Build your Own Search Service), which is an open API that enables developers to use Yahoo! Search to build search products leveraging their own data, content using Yahoo!'s crawling, indexing, and ranking algorithms. BOSS allows developers to submit queries via an API to retrieve web results in XML or JSON format. Yahoo does not impose limitations on usage and on daily queries. As opposed to Google's API, the API developed by Yahoo has liberal terms of use [10] and thus it is possible to use it in PESCaDO. Finally, Microsoft released the Bing API that enables the use of the Bing search engine and its results. Bing's search API provides search results programmatically as XML or JSON. However, Bing imposes some rather strict limitations regarding its use [11], such as the inability to modify and store the produced results.

An alternative to the aforementioned search engines are the meta search engines, which constitute information request services that use other search engines for their automated searches. More specifically, we consider as a meta-search engine, a system that forwards user queries to several search engines, aggregates the returned results, and presents the combined results to the user [5]. The most basic advantage of meta-search engines is the retrieval of results missed by a single primary engine. On the other hand, there are certain disadvantages of using the existing meta-search engines, such as the retrieval of limited results of each primary engine and the lack of advanced searching features. Some examples of meta-search engines are NorthernLight[13], MetaCrawler[14], DogPile[15],

[13] www.northernlight.com

[14] www.metacrawler.com

AltaVista [16]and MetaSpider [13]. To the best of our knowledge, none of the known meta-search engines provides an API.

3.3.1.2 KEYWORD SPICES

A methodology to generate automatically queries that result to domain specific results, is the use of keyword spices [6, 7]. The keyword spice method improves search performance by adding domain-specific keywords, called keyword spices, to the user's input query; the extended query is then forwarded to a general-purpose search engine. The extraction of the keyword spices is realized through a machine learning technique (i.e. decision-tree learning algorithm). The procedure for identifying the keyword spices comprises the manual selection of a sample of web pages that are classified into two classes *T* (relevant to the domain) or *F* (irrelevant to the domain). The pages are processed and nouns are extracted as keywords. Then, a decision tree learning algorithm is applied and a path that conjoins all keywords is created.

An alternative methodology to generate terms that are characteristic of a domain (i.e. keyword spice) is also described in [26]. In this work, a query formulation architecture, which employs the notion of context (i.e. domain) in order to automatically construct queries, is implemented. The proposed system uses semantic metadata extracted from the web page being consumed to automatically generate candidate queries. Therefore, the architecture presented supports query expansion with keyword spices, since the initial query inserted by the user, is expanded with words that characterize the context.

3.3.1.3 ANALYSIS OF SEARCH ENGINE RESULTS

Instead of employing keyword spices, another methodology is to perform analysis of the search engine results after the submission of empirically selected domain related queries. For instance, domain-related queries can be constructed from empirical knowledge of the specific domain, or using an ontology, a taxonomy or a lexicon.

In [13], the authors have demonstrated that web content mining clustering techniques involving Self-Organizing Map (SOM) algorithms can be applied to perform post-retrieval analysis of a retrieved document set, which generally improves the searching experience. A major drawback of post-retrieval analysis is the computation time and resources needed [5]. These limitations may be severe, especially for web-based search engines that have to handle thousands to millions of search queries per day. More specifically, Chen et al. created MetaSpider [13], which provides both meta-search and post-retrieval clustering functionality. The user can submit simple or more complex queries to MetaSpider, which retrieves a set of results that is further processed and finally outputs the pages that contain only the exact query phrase. Within the context of indentifying the content of the retrieved web pages, a tool was developed (Arizona Noun Phraser - AZNP) that indexes the key phrases that appear in each document retrieved by the Internet spiders. Finally, all the extracted phrases are sent to the SOM for automatic categorization.

An alternative methodology that is based on general-purpose search engines to build domain-specific ones is the filtering model. According to this method, the queries are forwarded to a general-purpose search engine and the irrelevant documents are sifted out. This is performed with the aid of domain-specific filters, which are based on empirically rules for identifying typical patterns found in the domain-specific web pages. This method was used by Shakes et al. [14] in creating Ahoy!, which is a search engine specialized for finding personal homepages. Ahoy! has a learning mechanism to assess the patterns of relevant URLs from previous successful searches, but overall accuracy basically depends on human knowledge.

[15] www.dogpile.com

[16] www.altavista.com

3.3.2 CRAWLING THE WEB

This methodology is based on employing crawlers for retrieving relevant content from the web. We consider two basic approaches, which are described in figures 3 and 4. In the first approach (fig 3), we consider a set of predefined qualified web sites and a general purpose crawler, while in the second (fig 4), we employ a domain focused crawler, which starts from some relevant seed points.

3.3.2.1 CRAWLING A PREDEFINED SET OF QUALIFIED WEB SITES

The most straightforward approach in order to build domain-specific web search engines is to collect and index only the relevant pages available on the web, thus limiting the scope of web crawlers to web sites with similar information. Architects of this type of search engines have the opportunity to thoroughly investigate each of the web sites included in the list of potentially interesting sites, and to develop tailor-made tools, which extend the search coverage of general search engines. Such specialized search engines can allow searching of the Invisible Web, which contains excluded web pages and database information currently not searchable by the major search engines [5]. However, the use of manually specified web page indices includes certain problems. The most important of them are the manual effort required to build and maintain the indices and the fact that it is not a scalable method that can catch up with the rapidly growing web [6, 7]. Another downside of this method is that the interfaces to the various sites to be searched can change frequently, so the developers have to continuously update their software to reflect these changes [6, 7]. Some examples of this technique are the LawCrawler[17], which searches for legal information on the Web and Moreover[18], which searches for the latest news from crawling.

In certain areas, people have collected and organized the most known sites in portals, in which domain-specific search engines search in databases generated by crawling a predefined set of qualified web sites [5]. The strategy of using of portals for obtaining domain specific information is usually followed by domain experts, who have the knowledge to how to formulate the search based on the scope and content of the database [1, 2, 4].

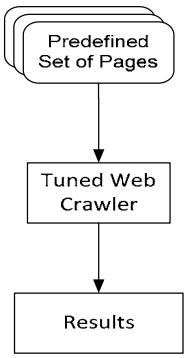


Figure 3.3: Crawling set of qualified web sites

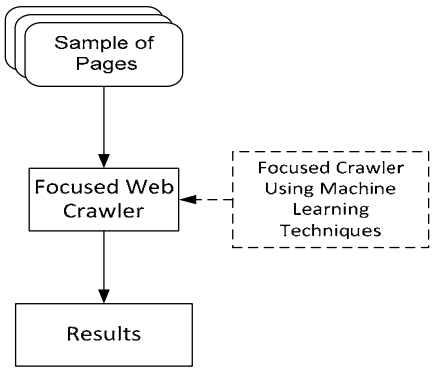


Figure 3.4: Using a focused crawler

Examples of Environmental Portals, Services and Sites:

[17] http://www.lawcrawler.com

[18] http:// www.moreover.com

- GEO Data Portal (http://geodata.grid.unep.ch/), authoritative source for data sets used by UNEP (United Nations Environment Programme)
- EcoEarth.Info - Environment Portal & Search Engine (http://www.ecoearth.info/)
- Environmental Industry, (http://www.cleanerproduction.com/directory/sectors/environm.htm)
- Pollen Reports for Europe (http://www.polleninfo.org/)
- COST ES 0602, Chemical Weather (http://www.chemicalweather.eu/Domains)

3.3.2.2 USING A FOCUSED CRAWLER

Another approach to develop a domain-specific web search engine is to employ web-crawling spiders that collect only relevant pages by using machine learning techniques [6, 7]. In this approach, a crawler (called focused crawler) is given a few starting nodes on the web, relevant to the topic of interest, and its goal is to seek out and collect other nodes that satisfy certain criteria related to the content of the source pages and the link structure of the web [5, 12]. Hence, when aiming to populate a domain-specific search engine, a web-crawling spider explores the Web in a directed fashion in order to find domain-relevant documents [3, 4]. These systems offer sophisticated search functions because they establish their own local databases and can apply various machine learning or knowledge representation techniques to the data but they are mostly suitable for those domains that have few web sites [6, 7]. Several approaches have been proposed for the implementation of a focused crawler.

McCallum et al. [3, 4] used reinforcement learning to build the domain-specific search engine, called Cora. The collection of new information was realized with the use of a spider that crawled the Web starting from the home pages of computer science departments and laboratories using reinforcement technique (i.e. learning optimal decision making from rewards or punishments). Then, an agent was trained off line, using collections of already found documents and hyperlinks and finally Naive Bayes classifiers were used to classify hyperlinks based on both the full text of the sources and anchor text on the links pointing to the targets [4, 12]. In another work, a focused crawler was implemented based on a machine learning approach [15], in which a supervised classifier was trained using positive and negative website examples. When the crawler fetched a new page, it was submitted to the classifier as a test case. If the classifier judged that the page was positive, outlinks from this page were added to the work pool as with standard crawlers, otherwise they were not considered for further crawling [5, 15]. In [12] a focused crawler for detecting documents relevant to the depressive illness is proposed. The method used was hypertext classification, which tries to classify documents by exploiting only the link and not the content information. The procedure followed, contained the use of a focused crawler starting from a set of relevant URLs (seed list), and using J48[19] in predicting future URLs [12]. Finally, in [16] a learnable focused crawling framework based on ontology was proposed. An Artificial Neural Network (ANN) was constructed using a domain-specific ontology and applied to the classification of web pages. The proposed crawling approach consists of three stages with distinct functions: the data preparation stage, the training stage, and the crawling stage. The data preparation stage is responsible for preparing training examples that are used for the ANN construction. In the training stage, an ANN is trained by the training examples through the use of a given domain-specific ontology that represents the background knowledge of crawling topics. Each web page in the training examples is pre-processed to get a list of entities. In the crawling stage, web pages are visited and the ANN determines whether or not they will be downloaded. Finally, the downloaded web pages are stored in a topic-specific web page repository.

[19] http://gautam.lis.illinois.edu/monkmiddleware/public/analytics/decisiontree.html

Examples of domain-specific search engines based on focused crawlers are the ResearchIndex, which is a specialized free search engine used for automatically finding Computer Science papers, the Deadliner for finding conference deadlines and the FlipDog[20] that looks at job postings at Web sites [5].

3.4 WEB SERVICES DISCOVERY

Web services (WS) are software systems designed to support interoperable machine-to-machine interaction over a network. Their interface is described in a machine-processable format (i.e. WSDL) and other systems interact with them in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards [17]. The basic platform of Web Services is XML plus HTTP. Web service technology has evolved around a stack of five technologies: Network (i.e. Communication between computers through a network), Transport (i.e. Sending and receiving of the messages), Packaging (i.e. Encodes messages in a standard format), Description (i.e. Describes a single service using WSDL), and Discovery (i.e. Provides a global directory of the existing –UDDI (Universal Description Discovery and Integration)) [18]. While SOAP is the standard protocol used in the packaging layer, XML-RPC and the recently emerging REST are also used.

At the early stages of service-oriented computing, finding relevant web services was mainly done by scanning through services registries (i.e. UDDI Business Registries). UDDI established the first uniform method that included details for integration of already existing systems and processes between business partners. A UDDI registry service is a WS that manages information about service providers, service implementations, and service metadata [20]. Although centralized registries, like UDDI, can provide effective methods for the discovery of Web services, they have problems associated with having centralized systems such as bottlenecks, problems related to providing quality of service measurements for registered Web services and thus they cannot provide any guarantee to the validity and quality of information it contains [19]. Other approaches focused on having multiple public/private registries grouped into registry federations such as METEOR-S. METEOR-S provides a discovery mechanism for publishing web services over a federated registry sources but, similar to the centralized registry environment, it does not provide any means for advanced search techniques, which are essential for locating appropriate business applications [19]. Issues related to the scalability of data replication, the increase of the number of web services disseminated throughout the web have driven researchers to find other alternatives. In order to address these issues, several simple web search engines have been implemented such as Binding point, Grand central and Web service list. These search engines provide only simple keyword search on web service descriptions based on the information provided by the WSDL files and the UDDI registry [22]. However, the keyword search paradigm is insufficient, because it fails to capture the underlying semantics of web services [22]. To address these challenges involved in searching for web services, Dong et al. [22] built Woogle, a web-service search engine, which involves doing unsupervised matching at the operation level, rather than supervised classification at the entire web service level research and thus the understanding of the operations in a web service is based on very limited amount of information.

Research is also conducted on investigating Web services on the Web. In [21], the authors provide an exploratory study on Web services on the Web. The study provides some details and statistics from Web services collected throughout the Web via Google API such as operation analysis, words distribution, etc. However, this study does not provide a complete view of Web services on the Web and focuses only on a single search engine. As an alternative, search engines such as Google, Yahoo, have become a new source for finding Web services. However, search engines do not recognize the significance for publishing service information on the Web in such a manner that meets the basic service properties (i.e. binding information, operations, ports, service endpoints, etc.) [19]. Finally, in [19] a targeted Web service crawler engine (WSCE) was developed that can be used for Web services discovery. WSCE actively crawls accessible UBRs and search engines to collect business and Web service information.

[20] www.flipdog.com

3.5 GRID DISCOVERY

The problem of the discovery of environmental service nodes in the Web is to a certain extent also common to modern open distributed computer systems, which offer various types of resources, abstracted as services across several organizational boundaries [30]. Most of them are based on grid and/or web services technologies and the resources are indexed to be searched for or to be discovered [28]. The quality and efficiency of the discovery decisively depends on the way data models and protocols describe the functionality and capacity of the given resources in the middleware. More recent discovery strategies emphasize the necessity to consider semantics in grid discovery [29]. Thus, to filter out the most suitable services, a semantic description may qualify all the characteristics of the service, and semantic discovery can make use of this information since it is capable of meaningful interactions.

A number of projects have been carried out in this field; cf. OntoGrid (FP6-511513) [31], MyGrid [32], FEARLUS-G (ESRC, RES-149-25-0011) [33], CombeChem [34], MediGrid Portal [35], Semantic Grid [36], etc. A few initiatives such as THREDDS [37] (Thematic Realtime Environmental Distributed Data Services) target specifically environmental services. More specifically THREDDS project is developing middleware to bridge the gap between data providers and data users in order to simplify the discovery and use of scientific data.

Most of the grid discovery approaches presuppose that the nodes to be searched post their functional "fingerprints", i.e., their functional capacity and coverage, for external inspection; see, e.g., OntoGrid and Semantic Grid. Thus, Semantic Grid relies on middleware that is already web service compliant. In the case of distributed environmental service nodes in the web – for instance, competing weather service nodes or complementary AQ service nodes – this cannot be assumed: as a rule, service nodes must be identified as being relevant and a profile of their capacity and functionality must be compiled.

3.6 STANDARDS

Currently, no specific standards exist for domain specific search engines and web service discovery. However, certain of the aforementioned approaches incorporate standards that provide to have uniform engineering or technical criteria, methods and practices. The most important of them which are related to web services and grid computing are summarized below.

SOAP [23]: It is a lightweight protocol intended for exchanging structured information in a decentralized, distributed environment. It uses XML technologies to define an extensible messaging framework providing a message construct that can be exchanged over a variety of underlying protocols.

WSDL [24]: It is an XML format for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information. The operations and messages are described abstractly, and then bound to a concrete network protocol and message format to define an endpoint.

OGSA (Open Grid Services Architecture) [25]: It describes an architecture for a service-oriented grid computing environment for business and scientific use, developed within the Global Grid Forum (GGF).

3.7 OPEN ISSUES

There are many open issues regarding the approaches that are related to the discovery of environmental service nodes in the Web. First, the methodologies for developing a web specific search engine seem to incorporate several drawbacks. In the case information is mined from a predefined list of sites, it is evident that although the information extracted is reliable, it is limited and constant updates of the crawler are required. On the other hand, the use of focused crawlers, while it is a more technologically advanced technique, it is far more time and computational power consuming. These disadvantages combined with the dispersal of information across many web sites, make the use of crawlers rather difficult. Furthermore, the use of existing search engines combined either with keyword spices

or analysis result techniques incorporates all the disadvantages of web search engines such as the inability to get results belonging to the Invisible Web. Finally, although web services provide well-defined and organized information, their discovery is based mainly on existing portals, where the developers register voluntary their services and thus only part of them can be acquired. Therefore, we can deduce that neither of the aforementioned techniques can fully cover the discovery of environmental nodes in the Web, when used separately. However, it is expected that an effective combination of the existing techniques can provide adequate coverage of the information dispersed in the Web. Apart of the combination of these techniques, research could be conducted towards the improvement of filtering techniques, the more advanced keyword spice extraction, as well as the development of more effective crawlers.

3.8 REFERENCES

[1] Suresh K. Bhavnani, “Domain-specific search strategies for the effective retrieval of healthcare and shopping information”, Conference on Human Factors in Computing Systems, Minnesota, USA, pp. 610 – 611, 2002.

[2] Nicole Mitsche, “Understanding the Information Search Process within a Tourism Domain-specific Search Engine”, Information and Communication Technologies in Tourism 2005, Springer Vienna, pp. 183-193, 2005.

[3] Andrew Mccallum, Kamal Nigam, Jason Rennie and Kristie Seymore, “A Machine Learning Approach to Building Domain-Specific Search Engines”, Proceedings of the 16th International Joint Conference on Artificial Intelligence, pp. 662 – 667, 1999.

[4] Andrew Mccallum, Kamal Nigam, Jason Rennie and Kristie Seymore, “Building Domain-Specific Search Engines with Machine Learning Techniques”, 1999.

[5] Wöber, K., "Domain Specific Search Engines", In Travel Destination Recommendation Systems: Behavioral Foundations and Applications, edited by D. R. Fesenmaier, H. Werthner, and K. Wöber, Cambridge, MA: CAB International, pp. 205-26, 2006.

[6] Satoshi Oyama, Takashi Kokubo, Toru Ishida, “Domain-Specific Web Search with Keyword Spices”, IEEE Transactions on Knowledge and Data Engineering, vol. 16 (1), pp. 17 – 27, 2004.

[7] S. Oyama T. Kokubo T. Ishida T. Yamada and Y. Kitamura, “Keyword Spices: A New Method for Building Domain-Specific Web Search Engines,” Proceedings of the 17th International Joint Conferences on Artificial Intelligence (IJCAI-01), pp. 1457-1463, 2001.

[8] Lance Whitney, “Bing grabs 10 percent of search market, ”http://news.cnet.com/8301-10805_3-10354394-75.html, September 2009.

[9] Terms of Use, Google AJAX Search API, *http://code.google.com/apis/ajaxsearch/terms.html*

[10] Terms of Use, Yahoo! Search BOSS Services Terms of Use, *http://info.yahoo.com/legal/us/yahoo/search/bosstos/bosstos-2317.html*

[11] Terms of Use, Bing Web Service Api, *http://www.bing.com/developers/tou.aspx*

[12] T. T. Tang, D. Hawking, N. Craswell, R. S. Sankaranarayana, “Focused crawling in depression portal search: A feasibility study”, Proceedings of the 9th Australasian Document Computing Symposium, Melbourne, Australia, December 13, 2004.

[13] Chen, H., Fan, H., Chau, M., Zeng, D., "MetaSpider: Meta-Searching and Categorization on the Web", Journal of the American Society for Information Science and Technology, Vol. 52, No 13, pp.1134-1147, Nov 2001.

[14] Jonathan Shakes, Marc Langheinrich and Oren Etzioni, “Dynamic reference sifting: a case study in the homepage domain”, In Proceedings of the 6th International World Wide Web Conference (WWW6), pp. 189– 200, 1997.

[15] Soumen Chakrabarti, Martin van den Berg, Byron Dom , “Focused crawling: a new approach to topic-specific Web resource discovery”, Computer Networks: The International Journal of Computer and Telecommunications Networking, Vol. 31 , Issue 11-16, pp. 1623 – 1640, May 1999.

[16] Hai-Tao Zhenga, Bo-Yeong Kanga and Hong-Gee Kim, "An ontology-based approach to learnable focused crawling", Information Sciences, Vol. 178, Issue 23, pp. 4512-4522, December 2008.

[17] Web Services Architecture, W3C Working group, February 2004, *http://www.w3.org/TR/ws-arch/*

[18] http://www.sitepoint.com/article/web-services-demystified, “Web Services Demystified”, Kevin Yank, 2002

[19] Al-Masri, E., and Mahmooud, Q.H.: Investigating Web Services on the World Wide Web, 17th International Conference on World Wide Web (WWW), Beijing, April 2008, pp. 795-804.

[20] Garofalakis, J., Panagis, Y., Sakkopoulos, E., Tsakalidis, A., "Web Service Discovery Mechanisms: Looking for a Needle in a Haystack?", International Workshop on Web Engineering, 2004

[21] Y. Li, Y. Liu, L. Zhang, G. Li, B. Xie, and J. Sun, “An Exploratory Study of Web Services on the Internet,” ICWS, pp. 380-387, 2007.

[22] X. Dong, A. Halevy, J. Madhavan, E. Nemes, and J. Zhang, “Similarity search for web services”. In: The International Journal on Very Large Databases, 2004.

[23] SOAP, W3C, April 2007, http://www.w3.org/TR/soap12-part1/

[24] WSDL, W3C, June 2007, http://www.w3.org/TR/wsdl20/

[25] OGSA, http://www.globus.org/ogsa/

[26] F. Menemenis, S. Papadopoulos, B. Bratu, S. Waddington, and Y. Kompatsiaris. "AQUAM: Automatic Query Formulation Architecture for Mobile Applications". In Proceedings of the 7th international Conference on Mobile and Ubiquitous Multimedia (Umea, Sweden, December 3 - 5, 2008). MUM '08, ACM, New York, NY.

[27] Tanenbaum, A.S. and M. Van Steen, "Distributed Systems:Principles and Paradigms", Prentice Hall, 2002.

[28] Taylor, I.J. and A. Harrison, "From P2P to web services and grids: Evolving Distributed Communities", Berlin: Springer, 2008.

[29] Pastore, S. "The necessity of semantic technologies in grid discovery", Journal of networks, 30(4):1-9. 2008.

[30] Tanenbaum, A.S. and M. Van Steen, "Distributed Systems:Principles and Paradigms", Prentice Hall, 2002.

[31] OntoGrid, http://www.ontogrid.net/ontogrid/index.html

[32] MyGrid, http://www.ebi.ac.uk/mygrid/

[33] FEARLUS-G, http://faceman.esc.abdn.ac.uk:8080/fearg/

[34] CombeChem, http://www.combechem.org/

[35] MediGrid Portal, http://www.medigrid.de

[36] Semantic Grid, www.semanticgrid.org

[37] THREDDS , *http://www.unidata.ucar.edu/projects/THREDDS/*

4 UNCERTAINTY METRICS DERIVATION

The uncertainty of measured or forecasted data and its propagation when the corresponding service forms part of an orchestrated service node configuration is a serious challenge.

4.1 METHODS TO EVALUATE UNCERTAINTY

A recent summary of the methods commonly used for uncertainty assessment and description in the environmental domain can be found in [Refsgaard et al., 2007][21] . They have chosen a total of 14 different methods to represent the commonly applied types of methods and tools:

Data uncertainty engine (DUE)

Error propagation equations

Expert elicitation

Extended peer review (review by stakeholders)

Inverse modelling (parameter estimation)

Inverse modelling (predictive uncertainty)

Monte Carlo analysis

Multiple model simulation

NUSAP

Quality assurance

Scenario analysis

Sensitivity analysis

Stakeholder involvement

Uncertainty matrix

For all of the methodologies more extensive descriptions and additional sources of information are available in (Refsgaard et al., 2007), here we re-present very shortly only the main findings of the extensive review to give an overview of the different methodologies and tools available for estimating the uncertainty of the environmental nodes.

4.1.1 DATA UNCERTAINTY ENGINE (DUE)

[21] Refsgaard, J.C. et al. 2007. Uncertainty in the environmental modelling process – A framework and guidance. Environmental Modelling & Software, 20:1543-1556.

Uncertainty in data may be described in 13 uncertainty categories depending on how data varies in time and space Each data category is associated with a range of uncertainty models, for which more specific probability density functions may be developed with different simplifying assumptions. Furthermore, correlation in time and space is characterised by correlogram / variogram functions. Data uncertainty is an important input when assessing uncertainty of model outputs. Assessment of data uncertainty is an area that theoretically is complex and full of pitfalls, especially when considering the correlation structure and its link with the scale of support.

4.1.2 Error propagation equations

The error propagation equations are widely used in the experimental and measurement sciences to estimate error propagation in calculations. The error propagation equations are valid only if the following conditions are met: (1) the uncertainties have Gaussian (normal) distributions; (2) the uncertainties for non-linear models are relatively small: the standard deviation divided by the mean value is less than 0.3; and (3) the uncertainties have no significant covariance. The method can be extended to allow non-Gaussian distributions and to allow for co-variances.

The main advantage of the error propagation equations is that they are easy and quick to use. The key limitations lie in the underlying assumptions that seldom hold, especially not for complex calculations. The error propagation equations are therefore mainly suitable for preliminary screening analysis.

4.1.3 Expert elicitation

Expert elicitation is a structured process to elicit subjective judgements from experts. It is widely used in quantitative risk analysis to quantify uncertainties in cases where there are no or too few direct empirical data available to infer on uncertainty. Expert elicitation typically involves the following steps: (1) Identify and select experts. (2) Explain to the expert the nature of the problem and the elicitation procedure. Create awareness of biases in subjective judgements and explore these. (3) Clearly define the quantity to be assessed and choose a scale and unit familiar to the expert. (4) Discuss the state of knowledge on the quantity at hand (strengths and weaknesses in available data, knowledge gaps, and qualitative uncertainties). (5) Elicit extremes of the distribution. (6) Assess these extremes: could the range be broader than stated? (7) Further elicit and specify the distribution. (8) Verify with the expert that the distribution that you constructed from the expert's responses correctly represents the expert's beliefs. (9) Decide whether or not to aggregate the distributions elicited from different experts .

Expert elicitation has the potential to make use of all available knowledge that cannot easily be formalised otherwise. The limitations are linked to the subjectivity of the results that are sensitive to the selection of experts. In case of differences among experts it may be difficult to safely quantify the uncertainties.

4.1.4 Extended peer review (review by stakeholders)

Extended peer review is the involvement of stakeholders in the quality assurance of the modelling process. Stakeholders' reasoning, observation and imagination are not bounded by scientific rationality. This can be beneficial when tackling ill-structured, complex problems. Consequently, the knowledge and perspectives of the stakeholders can bring in valuable new views on the problem and relevant information on that problem. The main strength of extended peer review is that it allows the use of extra knowledge from non-scientific sources. The key limitations lie in the difficulty for stakeholders to understand the sometimes complex and abstract concepts, to ensure representativeness of the selected stakeholders and in the power asymmetries that may be reproduced.

4.1.5 Inverse modelling (parameter estimation)

Parameter values are often estimated through inverse modelling. An optimal parameter set is sought by minimising an objective function, often defined as the summed squared deviation between the calibration targets (field data) and their simulated counterparts. Many software tools support inverse modelling and some universal

optimisation routines can be downloaded as freeware, e.g. PEST [22] and UCODE [23]Most inversion techniques have the benefit that they in addition to optimal parameter values also produce calibration statistics in terms of parameter-and observation sensitivities, parameter correlation and parameter uncertainties. An important limitation of these parameter uncertainty techniques is that the model calibration is based on a single model (with one possible model structure). Errors in the model structure will therefore wrongly be allocated to model parameter uncertainties. The estimated parameter uncertainties are thus uncertainties for the effective model parameter given both the model structure and available observations. This also means that estimated parameter uncertainties will not compensate adequately for the model structure uncertainty, when the model is used for prediction of conditions beyond the calibration base (e.g. when calibrating on groundwater flow and subsequently using the model to simulate solute transport).

4.1.6 INVERSE MODELLING (PREDICTIVE UNCERTAINTY)

In addition to parameter estimation some of the inverse optimisation routines include the ability to estimate predictive uncertainties. The method by which the predictive uncertainty is derived varies among the inversion routines. But common to many of the local optimisation routines based on non-linear regression, is that the prediction of interest is treated as an observation, and the regression algorithm is then used to quantify the effect of the parameter uncertainty on this ‘‘observation’’. Some methods rely on a semi-analytical solution in which the regression algorithm is used to compute either a predictive uncertainty interval for the output variable or uncertainty in the difference between a reference case and a scenario simulation. Other methods use the regression to seek the maximum or minimum value of the prediction under the constraint that the model must be calibrated at an acceptable level, which is defined by some predefined acceptance level of the objective function.

This method provides an objective estimate of the predictive uncertainty given the applied model structure. The main limitation, apart from assumptions on linearity and normally distributed residuals, is that uncertainty can only be predicted for data types for which observations exist. This means that uncertainties on variables that are interpolated or extrapolated compared to the available field data cannot be quantified by this method.

4.1.7 MONTE CARLO ANALYSIS

Monte Carlo Simulation is a statistical technique for stochastic model calculations and analysis of error propagation in calculations. Its purpose is to trace out the structure of the distributions of the model output. In its simplest form this distribution is mapped by calculating the deterministic results (realisations) for a large number of random draws from the individual distribution functions of input data and parameters of the model. As in random Monte Carlo sampling, pre-existing

Information about correlations between input variables can be incorporated. Monte Carlo analysis requires the analyst to specify probability distributions of all inputs and parameters, and the correlations between them. Both probability distributions and correlations are usually poorly known. Ignoring correlations and co-variance in input distributions may lead to substantial under-or over-estimation of uncertainty in model outcome. Advanced sampling methods have been designed such as Latin Hypercube sampling to reduce the required number of model runs needed to get sufficient information about the distribution in the outcome (mainly to save computation time).

A number of commercial and free software packages are available to do Monte Carlo analysis. In addition Monte Carlo functionality is built into many modelling software packages. The advantage of Monte Carlo analysis is its general applicability and that it does not impose many assumptions on probability distributions and correlations and

[22] http://www.pesthomepage.org/

[23] http://igwmc.mines.edu/freeware/ucode/.

that it can be linked to any model code. The key limitation is the large run times for computationally intensive models and the huge amount of outputs that are not always straightforward to analyse.

4.1.8 MULTIPLE MODEL SIMULATION

Multiple model simulation is a strategy to address uncertainty about model structure. Instead of doing an assessment using a single model, the assessment is carried out using different models of the same system. For instance, this can be realised by having alternative model codes with different process descriptions by having different conceptual models based on different geological interpretations.

The main advantages of this method are that the effects of alternative model structures can be analysed explicitly and that the robustness of the model predictions increases. An important limitation is that we cannot be sure whether we have adequately sampled the relevant space of plausible models and that important plausible model structures could be overlooked.

If the multiple model simulations is accompanied with quality assurance of each individual model (e.g. statistical comparison of model results vs. measurements (see 4.1.9 (3)) this method will provide not only a good estimate of the model uncertainties but also allows to rank the models based on the skill component which is most crucial for the end user. In the next chapter we present a practical example of utilizing this method for assessing the uncertainty and skills of a suite of regional scale air quality models.

4.1.9 QUALITY ASSURANCE

Quality assurance (QA) may be defined as protocols and guidelines to support the proper application of models. Important aims of QA are to ensure the use of best practise, to build consensus among the various actors involved in the modelling process and to ensure that the expected accuracy and model performance are in accordance with the project objectives.

Key elements of QA procedures include:

(1) framing of the problem and definition of the purpose of the modelling study;

(2) assessment of sources of uncertainties jointly by water manager, modeller and stakeholders and establishment of ac-curacy requirements by translation of the water manager and stakeholder needs to preliminary performance criteria;

(3) performance of model validation tests, i.e. testing of model performance against independent data that have not been used for calibration in order to assess the accuracy and credibility of the model simulations for situations comparable to those where it is intended to be used for; and (4) reviews carried out by independent auditors with subsequent consultation between the modeller, the water manager and possibly the stake-holders at different phases of the modelling project.

The HarmoniQuA [24]project has developed a comprehensive set of QA guidelines for multiple modelling domains combined with a supporting software tool, MoST.

QA improves the chances that best practise is used, it makes it possible to involve stakeholders into the modelling process in a formalised framework, and it improves the transparency and reproducibility. If not designed and performed thoroughly, QA may become a ‘rubber stamp’ and generate false credibility.

4.1.10 NUSAP

[24] http://www.harmoniqua.org

The NUSAP system for multidimensional uncertainty assessment aims to provide an analysis and diagnosis of uncertainty in science for policy. The basic idea is to qualify quantities by using the five qualifiers of the NUSAP acronym: numeral, unit, spread, assessment, and pedigree. NUSAP complements quantitative analysis (numeral, unit, spread) with expert judgement of reliability (assessment) and systematic multi-criteria evaluation of the different phases of production of a given knowledge base (pedigree). Pedigree criteria can be: proxy representation, empirical basis, methodological rigor, theoretical understanding, and degree of validation. NUSAP provides insight on two independent uncertainty-related properties expressed in numbers, namely spread and strength. Spread expresses inexactness whereas strength expresses the methodological and epistemological limitations of the underlying knowledge base. The two metrics can be combined in a Diagnostic Diagram, mapping strength of for instance model parameters and sensitivity of model outcome to spread in these model parameters. Neither spread alone nor strength alone is a sufficient measure for quality. Robustness of model output to parameter strength could be good even if parameter strength is low, if the spread in that parameter has a negligible effect on model outputs. In this situation our ignorance of the true value of the parameter has no immediate consequences. Alternatively, model outputs can be robust against parameter spread even if its relative contribution to the total spread in the model is high provided that parameter strength is also high. In the latter case, the uncertainty in the model outcome adequately reflects the inherent irreducible (stochastic) uncertainty in the system represented by the model. Uncertainty then is a property of the modelled system and does not stem from imperfect knowledge on that system. Mapping components of the knowledge base in a diagnostic diagram thus reveals the weakest spots and helps in setting priorities for improvement.

The strength of NUSAP is its integration of quantitative and qualitative uncertainty. It can be used on different levels of comprehensiveness: from a ‘back of the envelope’ sketch based on self elicitation to a comprehensive and sophisticated procedure involving structured, informed, in-depth group discussions on a parameter by parameter format. The key limitation is that the scoring of pedigree criteria is to a large extent based on subjective judgements. Therefore, outcomes may be sensitive to the selection of experts.

4.1.11 Scenario analysis

Scenario analysis aims to describe logical and internally consistent sequences of events to explore how the future may, could or should evolve from the past and present The future is inherently uncertain. Different alternative futures can be explored through scenario analysis. As such, scenario analysis is also a tool to deal explicitly with different assumptions about the future.

Scenarios can ensure that assumptions about future developments are made transparent and documented and are often the only way to deal with the unknown future. A limitation for qualitative scenarios is that it is difficult to test the underlying assumptions. For quantitative scenarios, the analysis is limited to those aspects of reality that can be quantified. Frequently, scenarios do not go beyond trend extrapolation and are surprise-free.

4.1.12 Sensitivity analysis

Sensitivity analysis (SA) is the study of how the variation in the output of a model can be qualitatively or quantitatively apportioned to different sources of variation, and of how the outputs of a given model depend upon the information fed into it.

Depending on the complexity of a model’s output space SA methods may range from the simple to the relatively complex. If a model’s output space is linear or approximates a hyperplane, SA may be conducted through a straightforward application of differential analysis. This is typically done by taking partial derivatives of the output with respect to one input, holding all other inputs constant. If a model’s output space is non-linear then the assumptions for differential analysis do not hold. Differential analysis may be conducted, but the analyst should be aware that the results may apply only to a narrow range of the output space. For this reason, differential analysis in this situation is referred to as Local SA.

The strength of SA is that it provides insight in the potential influence of all sorts of changes in input and helps discrimination across parameters according to their importance for the accuracy of the outcome. A limitation is the tendency of SA to yield an overload of information. Furthermore, SA most often takes the model structure and system boundaries for granted.

4.1.13 STAKEHOLDER INVOLVEMENT

Stakeholder involvement in not only the decision making process, but also in the modelling process, can help to assess and manage complex problems in a better way. This potential can be tapped in three ways: (1) by enabling them to articulate issues of concern and to improve the problem framing for re-search and policy; (2) by utilising their own knowledge and observations and their capacity to invent new options; and (3) by involving them actively in the quality control of the operational knowledge that is co-produced (ex-tended peer review, See section 4.4).

The key strengths of stakeholder involvement are that it in-creases the level of public accountability and it may increase the public support for implementation of subsequent management decisions.

4.1.14 UNCERTAINTY MATRIX

The uncertainty matrix can be used to identify and prioritise the most important uncertainties in a given model study. For a specific application the different sources of uncertainty are listed in the rows and the type of uncertainty associated to each source is noted and characterised. This may be done either quantitatively or qualitatively. The importance of each source may then be characterised by weighting depending on its impact on the modelling study in question. The sum of uncertainty may then be assessed, e.g. by use of the error propagation equations It may not be possible to identify all sources of uncertainty and/or assigning correct weightings from the project start. The matrix may thus be reassessed at each review, where new sources of uncertainty may be added or the weight of the uncertainty adjusted as more insight into

4.2 GEMS-RAQ MODEL SKILL EVALUATION

Project GEMS (Global and regional Earth-system (Atmosphere) Monitoring using Satellite and in-situ data) exploits advanced data assimilation of satellite and in-situ data in order to characterise the chemical composition of the atmosphere. The final aim was to develop an Integrated Forecasting System, which will be coupled to other chemical transport models and used to produce global forecasts of the chemical composition of the atmosphere. Here we concentrate on only one relevant outcome of the project , and describe briefly the adopted on-going evaluation methodology and the verification methods for the full suite of GEMS partner forecasts based on report by Agnew et, al, 2007.

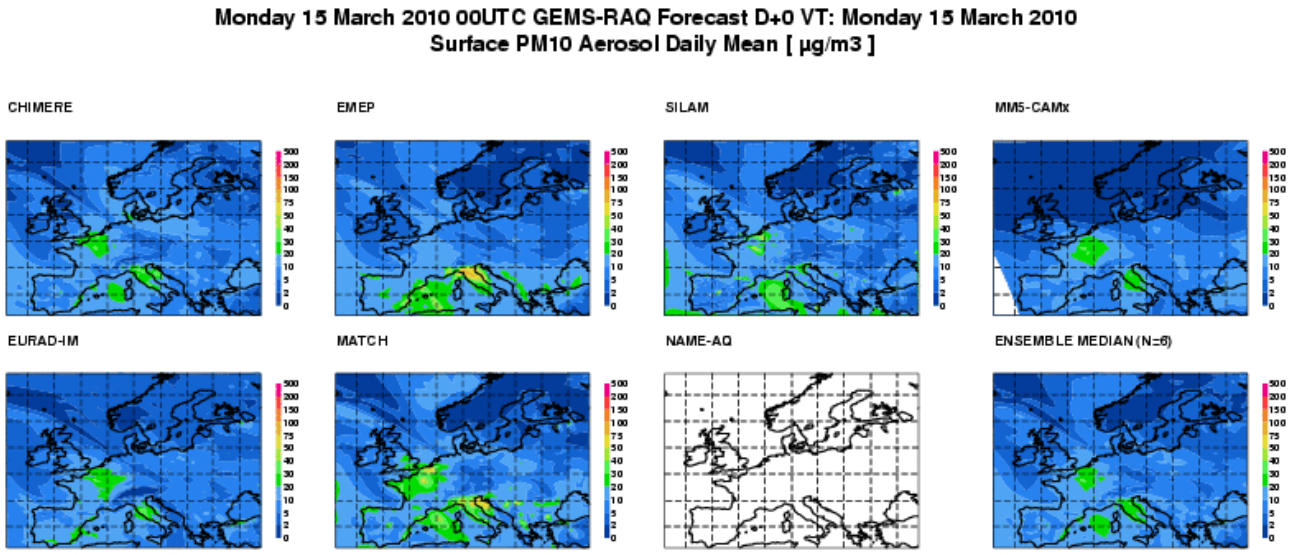


Figure 4.1: Forecasted European PM10 concentrations from: http://gems.ecmwf.int/d/products/raq/

4.2.1 MEASUREMENT DATA

When evaluating a model against measurement data, besides the definition of common statistical skill indicators, the question arises of selecting the "right" observation sites on which statistical indicators are to be evaluated. On one hand, air quality models compute the evolution of pollutant concentrations on grids; the concentrations can be thought of as averaged concentrations over the volume of each grid cell. On the other hand, observations are available from fixed measurement sites; they are local data and are influenced by local processes. These two values are different in nature and it is not straightforward to relate grid cell average concentrations to point concentrations.

Given an observation site, one has to answer the general question of spatial representativeness of observations, dependent on the characteristics of the site (topography, proximity to emission sources) and on the chemical species under evaluation (lifetime).

The selection of the "right" observation sites depends also on the model itself and on the purpose of the evaluation: do we need a model to produce an operational forecast or do we need a comprehensive model dedicated – for example - to the prediction of pollutant concentrations in emission reduction scenarios?

For the observation data to be usable by people that do not actually know the observation sites, a simple classification of these sites must be provided. As an example of such a classification for EuroAirnet, the European Air Quality monitoring network [25]developed since 1996 - is a selection of air quality monitoring stations in Europe, out of more than 6000 existing sites in about 30 countries. The selection is based on several criteria (EEA Technical Report No. 12, 1999), among which are i) the classification of monitoring stations ; ii) the area of representativeness of monitoring stations. These criteria were specified in order to provide a consistent set of monitoring stations across Europe. Table 4.1 illustrates the classification scheme of monitoring stations in EuroAirnet.

[25] http://air-climate.eionet.europa.eu/databases/EuroAirnet/index_html

Type of station	*Type of zone*	*Characterisation of zone*
Traffic (T)	Urban (U)	Residential (R)
Industrial (I)	Suburban (S)	Commercial (C)
Background (B)	Rural (R)	Industrial (I)
		Agricultural (A)
		Natural (N)
		Res/Com (RC)
		Com/Ind (CI)
		Ind/Res (IR)
		Res/Com/Ind (RCI)
		Agri/Natural (AN)

Table 4.1: Classification scheme of monitoring stations in EuroAirnet

For each monitoring station, an evaluation of its area of representativeness in terms of a radius is also provided. This is defined as the area in which the concentration does not differ from the concentration measured at the station by more than a specified amount. The area of representativeness varies with the station type. It depends on the concentration difference allowed in the definition and on the environment of the station, its morphology and sources. Determining the area of representativeness of a station requires either monitoring around the station or dispersion model calculations for the area in question and its surroundings. Such determinations are rarely performed. Thus the determination of station class is accompanied by an evaluation of the station's area of representativeness, taking into account the emission variations in the surroundings and any localized influence of sources further away, topographical features influencing the dispersion and transport of the emissions. Table 4.2 lists typical ranges of the area of representativeness (radius of area) for the various station types.

Station class	Radius of area
Traffic stations	*)
Industrial stations	10-100 m
Background stations:	
- Urban background stations	100m-1 km
- Near-city background stations	1-5 km
- Regional stations	25-150 km
- Remote stations	200-500 km

Table 4.2: Ranges of the area of representativeness for various station types

4.2.2 STATISTICAL MODEL VERIFICATION

Based on all available EuroAirnet observations the forecasting models are statistically verified. Although all partners /models are forecasting for the same domain, the resolution will be dependent on local model configurations and computing resources. No single statistical indicator is capable of capturing all aspects of model behavior thus a suitable verification strategy should produce a range of indicators which demonstrate the relative skill of the models to capture various features of the measured air quality.

An evaluation of some basic field statistics is described. These are useful in giving an overview of the performance of the model over the entire domain in terms of a simple bias, correlation and error statistic. However, a

problem arises in making a meaningful comparison of all the models on this basis since they will all run at different spatial resolutions. It is well known that the skill of a given run at low resolution may appear to be greater than that of a higher resolution simulation, due to the smoothing associated with the larger grid size. The higher resolution model is more susceptible to the 'double penalty' problem, whereby a local maximum is correctly forecast but in a slightly the wrong place, giving rise to two areas of error. However the lower resolution model would tend to show less skill in forecasting extremes and therefore it is important that any model assessment considers both performance aspects. Given forecast and observed values of a given species at site , it is possible to compute a variety of error statistics. Traditional metrics are normalized to include the mean bias and the root mean square error so that they provide a relative error. This is essential when comparing the bias of different chemical species which may be present in the atmosphere at very different concentration levels. The usual choice is to use the observed values for normalization, giving the normalized mean bias and normalized rmse

In comparing different GEMS RAQ models some may exhibit a trend to over-prediction and some to under-prediction. It is desirable to use a metric which treats both of these model deficiencies in a symmetric manner. A solution is to employ a normalization comprised of the arithmetic mean of the observed and forecast value, giving a modified mean bias, which gives a measure of forecast bias bounded by the values -2 to +2 and which performs symmetrically with respect to under and over-prediction. One weakness of this procedure is that the normalization employed varies slightly between different forecast models due to the dependence on forecast values. However, in most cases the variation is not expected to significantly distort the calculated bias.

The traditional rationale for employing the (normalized) rmse as an indicator of overall forecast error is two fold: (i) by squaring the errors before combining, this measure removes any cancellation of under and over-prediction; (ii) in cases where the spread of errors approximates to a well-known distribution the rmse can be attributed with a physical significance. In the present case the errors are not expected to conform to any well-known distribution. In addition, the rmse suffers from similar deficiencies as the mean bias, displaying an asymmetry with respect to under and over forecasting. A further issue is that the rmse gives added weight to those errors having greater magnitude, as a consequence of the squaring operation. In view of these issues the fractional gross error is used as the indicator of overall forecast error , which is essentially a relative version of the commonly used 'mean absolute error'. The modified mean bias indicates the extent to which the model under or over-predicts the set of observations, whilst the fractional gross error gives a measure of the overall forecast error. An additional metric proposed for comparing forecast and observation fields over the whole GEMS domain is the correlation coefficient. This is needed to indicate the extent to which patterns in the forecast match those in the observations.

The statistical measures discussed above, when taken together provide a valuable indication of model performance over the entire spatial domain. The results are expressed numerically and a time series of each quantity (bias, gross error, correlation) can be plotted to show trends over time. However Taylor (2001) has shown that a single diagram can be used to summarize the basic statistical measures of pattern rmse and correlation and thus display the relative performance of a number of forecast models in a visually accessible manner. A reference field, such as an analysis field or in our case the set of observations, is normally plotted along the x-axis. Normalization of the forecast and observation fields via the observation standard deviation places the reference point at a value of unity along the x-axis. The radial distance from this point to that of the forecast field gives the pattern rmse. Cf. Figure 4.2.

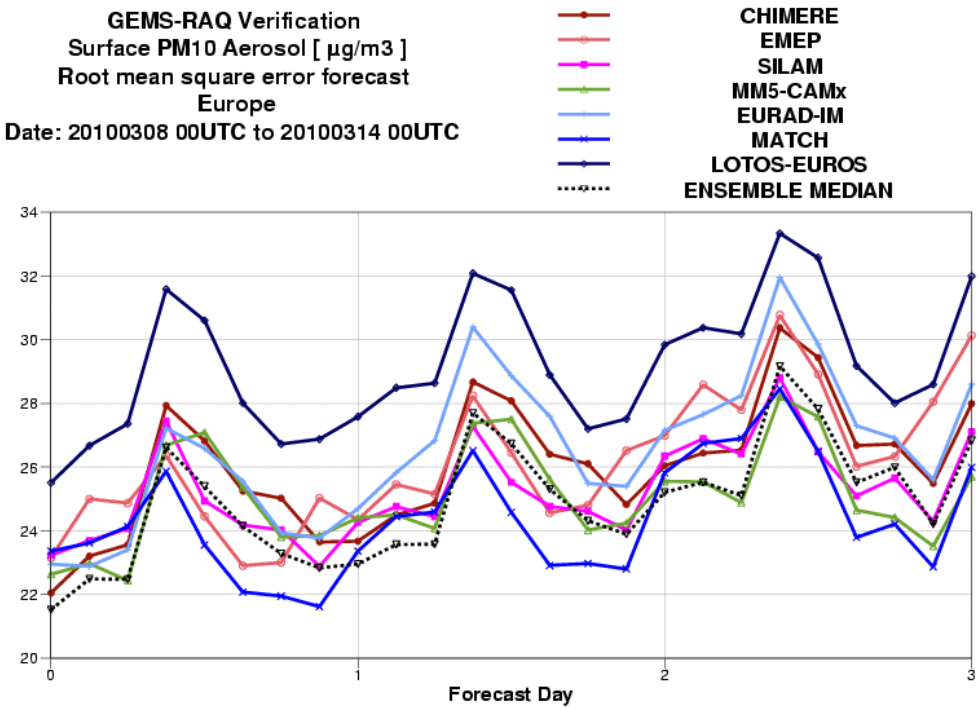


Figure 4.2: RMSE for PM10 forecasts. @15/03/2010 [26]

The verification measures described above provide information about the forecast errors under all conditions, regardless of the magnitude of pollutant concentration. We need however also metrics which provide information regarding forecast skill specifically at those times when pollutant levels are elevated and pose a greater risk to human health. For a given chemical species threshold levels have been identified for which it is recommended that the public be either informed or given warnings regarding possible adverse health effects. For example for ozone an information threshold of 180µgm-3 and a warning threshold of 240µgm-3 have been specified. It is of interest to assess the skill that models possess in predicting exceedance of given thresholds. The method traditionally employed in this case is to construct a 2x2 contingency table simply summarizing how many correctly forecasted events vs. false alarms were observed. Based on this table, a metric which is easy to calculate, not sensitive to chosen threshold or base rate and also robust is the odds ratio skill score The 'odds' is defined as the ratio of probability that an event (such as the exceedance of a threshold) occurs, to probability that it does not occur. It is a non-negative number with a value greater than unity when a success (good forecast) is more likely than a failure (bad forecast). The odds ratio is calculated on the assumption that forecast skill can be judged by comparing the odds of a success (a hit) to the odds of failure (false alarm): This score can further be transformed to final, easier to understand skill score (ORSS) ranging from -1 to +1 . The exact equations and definition are all given in the original reference [Agnew et al., 2007].

4.3 SUMMARY

The uncertainty of measured or forecasted data and its propagation when the corresponding service forms part of an orchestrated service node configuration is a serious challenge. We presented a short overview of different methods applicable for practical uncertainty assessment based on the recent review by Refsgaard et al. [2007] , and as a practical example presented a description of an operative system for evaluating Regional Air Quality (RAQ) forecasts which were developed and utilized in GEMS-RAQ-subproject [Agnew et al, 2007]. Basic statistical metrics for assessing (forecast-observation) bias, error and correlation are described. The ability of RAQ models to forecast

[26] http://gems.ecmwf.int/d/products/raq/

exceedance events is a key performance metric and a skill score based on the odds ratio is proposed for this case. This measure is not dependent on the event base rate, is not sensitive to the magnitude of the thresholds chosen and has a number of other properties which make it suitable for the evaluation of RAQ forecasts. Another crucial aspect of forecast verification is the choice of measurements sites. Some key issues related to this topic were discussed.

5 PROTOCOLS FOR CONNECTING ENVIRONMENTAL SERVICES

A considerable effort has been undertaken by steering organizations, funding institutions and the research community to push forward the standardization of protocols for exchange of data and knowledge in environmental applications as well as the standardization of web service definitions. Data protocols have been addressed within several initiatives of OGC (Open Geospatial Consortium) [OGC, 2010]. Other notable organizations are W3C (the World Wide Web Consortium) [W3C, 2010] and OASIS (Organization for the Advancement of Structured Information Standards) [OASIS, 2010].

The subsections 5.1 to 5.3 describe standards for connecting generic services, finding and securing them. Implementations of these standards are the building blocks for a service oriented architecture (SOA). The following sections cover standards for geographic formats and services which are relevant for the environmental application of PESCaDO.

5.1 WEB SERVICES

There are several competing approaches to Web Service development. Web Services architectures can be broadly distinguished between 'resource-oriented' and 'service-oriented' [Snell, 2004].

Acronym	Description
SOAP Simple Object Access Protocol	SOAP (Simple Object Access Protocol) is a protocol specification that defines a uniform way of passing XML-encoded data. In also defines a way to perform remote procedure calls (RPCs) using HTTP as the underlying communication protocol. SOAP arose from the realization that no matter what the current middleware offerings are, they need a Wide Area Network wrapper. Architecturally, sending messages as plain XML has advantages in terms of ensuring interoperability. The middle-ware players seem willing to put up with the costs of parsing and serializing XML in order to scale their approach to wider networks. Submitted in 2000 to the W3C as a Note by IBM, Microsoft, UserLand, and DevelopMentor, the further development of SOAP is now in the care of the W3C's XML Protocols Working Group (Web Service Activity).
UDDI Universal Description, Discovery and Integration Service	UDDI (Universal Description, Discovery and Integration Service) provides a mechanism for clients to dynamically find other Web services. Using a UDDI interface, businesses can dynamically connect to services provided by external business partners. A UDDI registry is similar to a CORBA trader, or it can be thought of as a DNS service for business applications. A UDDI registry has two kinds of clients: businesses that want to publish a service (and its usage interfaces), and clients who want to obtain services of a certain kind and bind programmatically to them.
WSDL	WSDL (Web Services Definition Language) provides a way for service providers to describe the basic format of Web service requests

Web Services Definition Language	over different protocols or encodings. WSDL is used to de-scribe what a Web service can do, where it resides, and how to invoke it. While the claim of SOAP/HTTP independence is made in various specifications, WSDL makes the most sense if it assumes SOAP/HTTP/MIME as the remote object invocation mechanism. UDDI registries describe numerous aspects of Web services, including the binding details of the service. WSDL fits into the subset of a UDDI service description. WSDL defines services as collections of network endpoints or ports. In WSDL, the abstract definition of endpoints and messages is separated from their concrete network deployment or data format bindings. This allows the reuse of abstract definitions of messages, which are abstract descriptions of the data being exchanged, and port types, which are abstract collections of operations. The concrete protocol and data format specifications for a particular port type constitute a reusable binding. A port is defined by associating a network address with a reusable binding; a collection of ports define a service.
REST Representational state transfer	Representational state transfer (REST) was first introduced in 2000 as an „architectural style for distributed hypermedia systems“ by [Fielding, 2000]. REST components perform actions on a resource by using a representation to capture the current or intended state of that resource and transferring that representation between components. RESTful Web Services are typically built on some kind of Resource-Oriented Architecture which adheres to the principles and constraints of the REST architectural style. The concept of 'Resources' is the central most important concept of REST: It revolves around use of a very limited set of operations (the 'Uniform Interface') to a (potentially very large) set of data endpoints (resources). Therefore RESTful Web Services are classified as 'resource-oriented'. On the implementation level, RESTful Web Services use well understood and accepted internet technologies and standards (e.g. HTTP, HTML, XML, etc.). The central idea behind RESTful Web Services is that - instead of building additional complex layers of functionality on top of existing Web technologies - the standards, technologies and underlying principles that made the WWW scalable, usable, accessible and flexible can (and should) be reused for Web Service implementation [Bommersbach, 2009].

5.2 SEMANTIC WEB SERVICES

For the connection of environmental service nodes at the semantic level, especially the W3C-promoted SAWSDL recommendation concerning the semantic annotation of the Web Service Description Language (WSDL) and XML schema documents, which has been supported by the WS2 Project (IST-FP6-004308)27, is of relevance:

Acronym	Description
SAWSDL Semantic Annotations for WSDL and XML Schema	Semantic Annotations for WSDL and XML Schema (SAWSDL) defines how to add semantic annotations to various parts of a WSDL document such as input and output message structures, interfaces and operations. The extension attributes defined in this specification fit within all versions of WSDL and XML Schema extensibility frameworks. For example, this specification defines a way to annotate WSDL interfaces and operations with categorization information that can be used to publish a Web service in a registry. The annotations on schema types can be used during Web service discovery and composition [SAWSDL, 2007].

5.3 SECURITY STANDARDS

Acronym	Description
SAML Security Assertion Markup Language	The Security Assertion Markup Language (SAML) is a product of the OASIS Security Services Technical Committee. It is an XML-based standard for exchanging authentication and authorization data between an identity provider and a service provider (security domains). The main problem that SAML is trying to solve is the Single Sign-On (SSO). To facilitate that SAML assumes the principal (a user of a service) has enrolled with at least one identity provider. This identity provider is expected to provide authentication services to the principal. SAML does not specify the implementation of these services. Thus a service provider relies on the identity provider to identify the principal. At the principal's request, the identity provider passes a SAML assertion to the service provider. On the basis of this assertion, the service provider makes an access control decision.
XACML eXtensible Access Control Markup Language	The eXtensible Access Control Markup Language (XACML) allows administrators to define the access control requirements for their application resources. It is a declarative access control policy language implemented in XML and a processing model, describing how to interpret the policies. XACML is an outgrowth of work to support SAML's very basic authorization decision query protocol, although XACML is not intended to be limited to use with SAML protocols. Latest version 2.0 was ratified by OASIS standards organization on 1 February 2005. The planned version 3.0 will add generic attribute

[27] http://www.w3.org/2004/WS2/

	categories for the evaluation context and policy delegation profile (administrative policy profile) [XACML, 2005].
GeoXACML	GeoXACML is an extension to the OASIS XACML standard, which has been approved by the OGC. The primary goal of the GeoXACML extension is to support combinations of class-based, object-based and spatial permissions.

5.4 GEO FORMATS

Acronym	Description
GML Geography Markup Language	The Geography Markup Language (GML) is an XML encoding for the transport and storage of geographic information, including both the geometry and properties of geographic features. This specification defines the XML Schema syntax, mechanisms, and conventions that: Provide an open, vendor-neutral framework for the definition of geospatial application schemas and objects; Allow profiles that support proper subsets of GML framework descriptive capabilities; Support the description of geospatial application schemas for specialized domains and information communities; Enable the creation and maintenance of linked geographic application schemas and datasets; Support the storage and transport of application schemas and data sets; Increase the ability of organizations to share geographic application schemas and the information they describe.
KML Keyhole Markup Language	KML is an XML language focused on geographic visualization, including annotation of maps and images. Geographic visualization includes not only the presentation of graphical data on the globe, but also the control of the user's navigation in the sense of where to go and where to look. [KML, 2010] Originally developed by Google, KML version 2.2 has been approved by the OGC as an Open Standard (OGC KML).

5.5

5.6 GEOGRAPHIC SERVICES

Acronym	Description
WCS Web Coverage Service	The Web Coverage Service (WCS) supports electronic interchange of geospatial data as "coverages" – that is, digital geospatial information representing space-varying phenomena. This document specifies how a Web Coverage Service (WCS) serves to describe, request, and deliver multi-dimensional coverage data over the World Wide Web. This version of the Web Coverage Service is limited to describing and requesting grid (or "simple") coverages with homogeneous range sets.
WFS Web Feature Service	The purpose of the Web Feature Server Interface Specification (WFS) is to describe data manipulation operations on OpenGIS Simple Features (feature instances) such that servers and clients can "communicate" at the feature level. Therefore, a Web Feature Server request - like those supported in many GIS and RDBMS packages - consists of a description of the query and data transformation operations that are to be applied to WFS enabled spatial data warehouses on the Web. The request is generated on the client and is posted to a WFS server. The WFS Server "reads" and executes the request returned in a feature set as GML. A GML enabled client then can use the feature set. A WFS client implementation supports the dynamic exploitation and access of feature data and associated attributes on the web from any server product that implements WFS. This capability opens the door to enhanced spatial analysis, modeling and other operations based on the intelligence of the attributed data. Beyond feature access, there is an additional set of interfaces in the WFS for supporting simple transactions: Create a Feature, Delete a feature, and Update a feature.

WMS Web Map Service	This International Standard specifies the behavior of a service that produces spatially referenced maps dynamically from geographic information. It specifies operations to retrieve a description of the maps offered by a server, to retrieve a map, and to query a server about features displayed on a map. This International Standard is applicable to pictorial renderings of maps in a graphical format; it is not applicable to retrieval of actual feature data or coverage data values. A Web Map Service (WMS) produces maps of spatially referenced data dynamically from geographic information. This International Standard defines a "map" to be a portrayal of geographic information as a digital image file suitable for display on a computer screen. A map is not the data itself. WMS-produced maps are generally rendered in a pictorial format such as PNG, GIF or JPEG, or occasionally as vector-based graphical elements in Scalable Vector Graphics (SVG) or Web Computer Graphics Metafile (WebCGM) formats.
SLD Styled Layer Descriptor	The SLD is an encoding for how the Web Map Server (WMS 1.0 & 1.1) specification can be extended to allow user-defined symbolization of feature data. This document addresses the need for geospatial consumers (either humans or machines) to control the visual portrayal of the data with which they work. The ability for a human or machine client to define these rules requires a styling language that the client and server can both understand. SLD can be used to portray the output of Web Map Servers, Web Feature Servers and Web Coverage Servers.

5.7 SENSOR WEB ENABLEMENT (SWE)

Accessing sensors is not in the focus of PESCaDO. Nevertheless, the topic of accessing sensor information via the web is currently the biggest driving force in the field of environmental systems.

The Sensor Web Enablement (SWE) initiative run by the OGC builds a framework of open standards for exploiting Web-connected sensors and sensor systems of all types: flood gauges, air pollution monitors, stress gauges on bridges, mobile heart monitors, Webcams, satellite-borne earth imaging devices and countless other sensors and sensor systems. Examples of important standards coming out of SWE are:

Observations & Measurements (O&M) - The OGC Observations and Measurements Encoding Standard (O&M) defines an abstract model and an XML scheme encoding for sensor observations [Cox(ed.), 2007].

Sensor Model Language (SensorML) - SensorML is an OGC standard markup language (using XML scheme) for providing descriptions of sensor systems. By design it supports a wide range of sensors [Botts/Robin, 2007].

Sensor Observation Service (SOS) - The OGC Sensor Observation Service specifies an API for accessing deployed sensors and retrieving observation data (results). The goal of SOS is to provide access to observations from sensors and sensor systems in a standard way that is consistent for all sensor systems including remote, in-situ, fixed and mobile sensors [Na/Priest(ed.), 2007].

5.8 REFERENCES

[Bommersbach, 2009]: Ralf Bommersbach, Specification and Implementation of a discoverable RESTful Web Service for Sensor Observations, 2009

[Botts/Robin, 2007]: Mike Botts (ed.), Alexandre Robin (ed.), Sensor Model Language (SensorML), OpenGIS Implementation Specification, Version 1.0.0 (2007), Open Geospatial Consortium Document #07-000

[Cox(ed.), 2007]: Simon Cox (ed.), Observations and Measurements Part 1 - Observation schema, OpenGIS Implementation Standard, Version 1.0 (2007), Open Geospatial Consortium Document #07-022r1

[Fielding, 2000]: Roy Thomas Fielding, Architectural Styles and the Design of Network-based Software Architectures, 2000

[KML, 2010]: KML, http://www.opengeospatial.org/standards/kml/

[Na/Priest(ed.), 2007]: Arthur Na (ed.), Mark Priest (ed.), Sensor Observation Service, OpenGIS Implementation Standard, Version 1.0 (2007), Open Geospatial Consortium Document #06-009r6

[OGC, 2010]: Open Geospatial Consortium, 2010, http://www.opengeospatial.org/

[OASIS, 2010]: Organization for the Advancement of Structured Information Standards, http://www.oasis-open.org/

[SAWSDL, 2007]: Joel Farrell, Holger Lausen (ed.), Semantic Annotations for WSDL and XML Schema, 2007

[Snell, 2004]: James Snell, Resource-oriented vs. activity-oriented Web services, 2004,

http://www.ibm.com/developerworks/xml/library/ws-restvsoap/

[W3C, 2010]: World Wide Web Consortium, 2010, http://www.w3.org/

[XACML, 2005]: Tim Moses (ed.), eXtensible Access Control Markup Language (XACML) Version 2.0, 2005

6 ENVIRONMENTAL NODE ORCHESTRATION

6.1 OVERVIEW

Computer-based orchestration of environmental services is not new. The results of meteorological data measuring networks are chained-in into meteorological forecast services. Meteorological forecast services are chained-in into AQ-services. Flood and lightning warning models are increasingly automatically fed with the output of meteorological services and measured and forecasted data from a variety of different environmental services are fed into risk assessment models. As a rule, in these applications, the orchestration is done manually, with proprietary interface functions and hard-wired connections. More recently, applications with a flexible connection of distributed environmental services have been created (as in the projects GEMS, PROMOTE, ORCHESTRA, SANY, etc. and in the large scale initiatives INSPIRE, GMES and GEOSS). However, in these recent applications, the dynamic selection and connection of nodes based on their quality and content, judged relatively to the other nodes, plays a minor or no role at all.

6.2 ORCHESTRATION OF CHEMICAL WEATHER FORECAST MODELS IN EUROPE

Methods that include a combination of weather forecasting and atmospheric chemistry simulations are here referred to as chemical weather forecasting. The definition of chemical weather therefore extends the concept of air

quality forecasting. For instance, air quality forecasting models using statistical methods that are not based on weather forecasting models, are therefore by definition not chemical weather forecasting models. Chemical weather forecasting requires access to meteorological and air pollutant concentration fields and measurements emission inventories and physiographic data. The users of chemical weather or air quality forecasts include European citizens, public authorities and agencies that are in charge of environmental impact assessments and public health.

There are prominent ongoing European projects in this area, in particular within the EU-ESA (European Space Agency) programme GMES (Global Monitoring for Environment and Security, http://www.europa.eu.int/comm/space/gmes/index_en.html), such as GEMS (http://gems.ecmwf.int/) and PROMOTE (PROtocol MOniToring for the GMES Service Element, (http://www.gse-promote.org/) , Poupkou et al. 2006). The GMES Atmospheric Services focus on operational monitoring and forecasting of atmospheric composition, dynamics and thermodynamics through advanced exploitation of satellite and in-situ data, on a European, national and local level. Clearly, currently there are also several other related EU-funded projects, such as MEGAPOLI (Baklanov et al. 2008), CITYZEN, EUCAARI (http://www.atm.helsinki.fi/eucaari/) and EUSAAR (http://www.eusaar.net/). Within the GEMS project, analyses and 72h forecasts have been presented using ten state-of-the-art regional air quality models from nine countries (BOLCHEM, CAC, CHIMERE, EMEP, EURAD, MATCH, MOCAGE, MM5-UAMV, NAME and SILAM) on a quasi-operational daily basis (http://www.ecmwf.int/). The models rely on the operational meteorological forecasts of the European Centre for Medium-Range Weather Forecasts, as well as on GEMS global chemical weather data. They all consider the same high-resolution (~ 8 km) anthropogenic and biogenic emissions inventories.

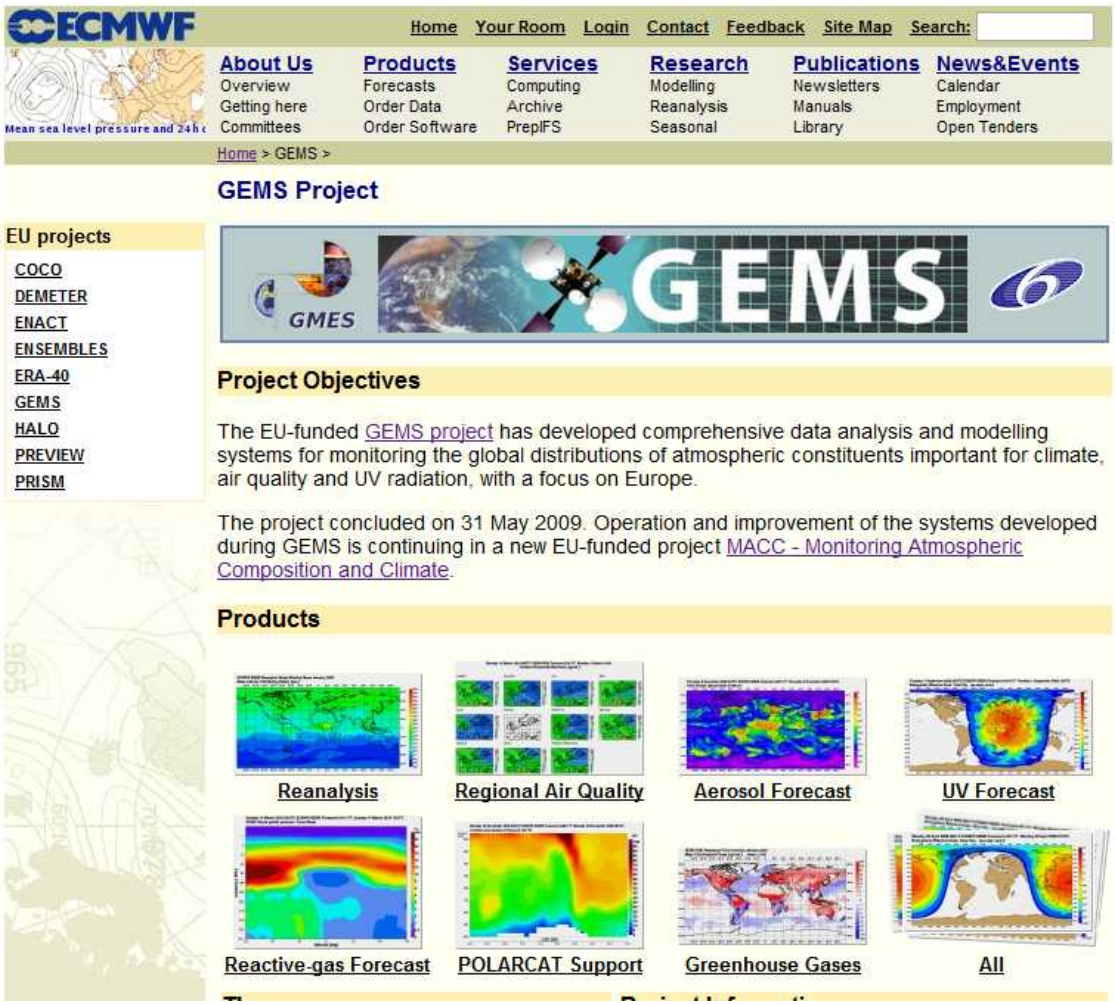


Figure 6.1: GEMS Project portal. (http://gems.ecmwf.int/)

An example of a small-scale network of operational air quality services has been constructed within the first and second stages of the PROMOTE project. These projects, however, have a selected membership and are development-oriented; these cannot therefore involve all stake-holders in a comprehensive way, such as the national environmental agencies.

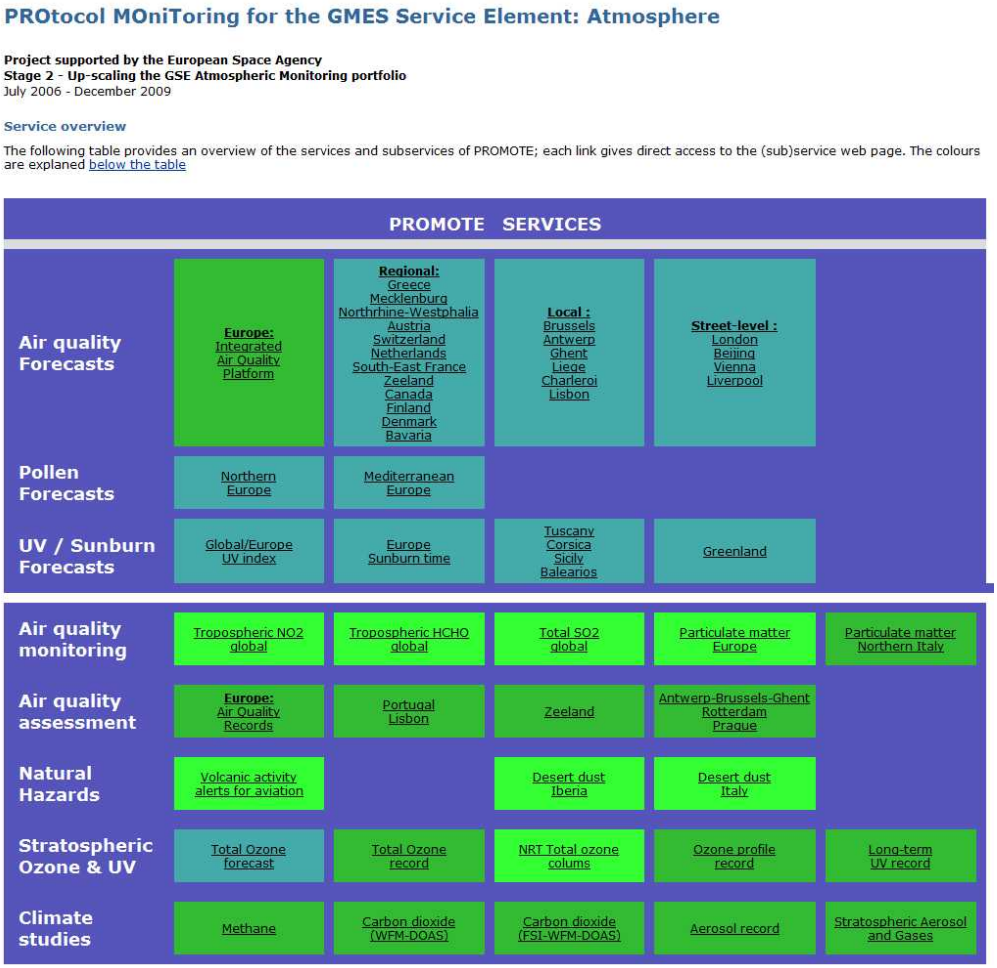


Figure 6.2: Promote Services. (http://www.gse-promote.org/index2.html)

The PROMOTE -portal provides links to air quality forecasts (including an ensemble forecast), observed values and also to archived environmental data.

Another relevant activity is the Global Earth Observation and Monitoring (GEOmon) project (www.geomon.eu), the goal of which is to sustain and analyze European ground based observations of atmospheric composition and their complementarity with satellite observations. It aims to lay the foundation for a European contribution to GEOSS (Global Earth Observation System of Systems, *http://www.epa.gov/geoss/*) and to optimize the European strategy of environmental monitoring in the field of atmospheric composition measurements. Key deliverables include the provision of access to data through a common data centre (for atmospheric composition data), and the dissemination of data and data products through user friendly tools.

However, currently there is no comprehensive review of available CWF services in the internet, for the various European regions and urban areas. Even for expert users, it is therefore far from evident to find out, e.g., (i) which of these services would be best applicable for any specific requirements, (ii) which services are available for a specific

geographic domain, and (iii) how accurate, and reliable are the forecasts of any given system. For example, various CWF systems may differ substantially regarding, the anthropogenic or natural sources of particulate matter that are included in the computations.

To answer at least some of these open questions an European chemical weather forecasting portal that includes links to a substantial number of CW forecasts in Europe in a user-friendly graphical format (*http://www.chemicalweather.eu/Domains*) has been recently developed. This interactive application is a public domain portal, in which the forecasts can easily be viewed, accessed and also added. The system is continuously updated, to incorporate various related information and services. All selected models are CWF models, i.e., these include a combination of numerical weather forecasting and atmospheric chemistry modeling. The portal includes primarily CWF models on European and regional scales; although some urban, local and global scale models and model applications have also been included. Such a single point of reference for the European CWF systems has not been operational previously. The portal differs from the above mentioned GMES portals by providing a broader membership across different communities. It also includes CW forecasts for more diverse domains from global and European scales to urban scale, whereas the GEMS portal solely uses a European domain and the PROMOTE portal provides forecasts mainly on national and urban scales.

Figure 6.3: Cost ES0602, Chemical Weather portal (http://www.chemicalweather.eu/Domains)

The user can select an area on the European map (see *Figure 6.3*) and will automatically receive a list of the available air quality forecasting systems that contain chemical weather predictions for the that area. Additional information provided include direct links to the forecasting systems covering the location of interest, as well as the scientific group that has developed and is maintaining it, the model(s) applied, a short description of the system, and a link to related records within the Model Documentation System of the European Environment Agency.

6.3 SUMMARY

We have provided short description of some state-of-art Air Quality (AQ) information portals operating in Europe. Although, they cover already quite well most of the available operative AQ services in Europe, little if any effort has been put in any automated orchestration of the multitude of separate services. Also, from the presented portals and services only the GEMS-RAQ portal is at the moment providing continuous online information on the quality of the included forecasting services, so the task of selecting the most reliable and suitable service is still largely left as the sole responsibility of the end-user.

7 ENVIRONMENTAL ONTOLOGY ALIGNMENT AND EXTENSION

PESCaDO's overall technological goal is the development of an operational workbench for the orchestration of environmental services and multilingual delivery of their output. This workbench, and the techniques supporting it to be developed in PESCaDO, will be founded on the availability of an ontological representation of the environmental domain, in order to guarantee semantic orchestration of heterogeneous environmental service nodes, decision support and environmental information production.

This ontological representation has to be as much as possible comprehensive, in order to cover adequately the environmental domain and, in particular, PESCaDO use cases domains. Although such ontology is not currently available among the state of the art resources, it is unlikely that it will have to be built from scratch. In fact, several ontologies covering (specific) parts of the environmental domain (or environmental-related domains) have been already proposed, and are accessible. In certain cases, the domains described by some of these ontologies may overlap, and hence some entities may be described in more than one ontology. Furthermore, parts of the knowledge of the environmental domain relevant for PESCaDO may not be covered by any available ontology.

In view of this, the ontology construction task in PESCaDO will have to undergo two specific activities:

- the discovering of candidate semantic correspondences between different ontologies (*ontology alignment*)
- the extension of existing ontologies with new concepts and relation that emerges from the automatic or manual analysis of data (*ontology extension*).

The ontology alignment and ontology extension problems have been already deeply investigated in the literature, and several tools and techniques to deal with these problems are available. We briefly present some of the main results below, after introducing a brief overview of the state of the art in terms of ontological resources for the environmental domain.

7.1 GLOSSARY/ABBREVIATION

OWL	Web Ontology Language – a languages for authoring ontologies
RDF	Resource Description Framework – a standard model for data interchange on the Web.

SPARQL	A recursive acronym for 'SPARQL Protocol and RDF Query Language' – SPARQL is a query language for RDF graphs
URI	Uniform Resource Identifier – a string of characters used to identify a name or a resource on the Internet.

7.2 Overview of Existing Environmental Ontologies

The state of the art on environmental OWL ontologies abounds of candidates which may be relevant for domains of interest of PESCaDO. We briefly present some prominent candidates, while a more detail analysis and comparison will be carried out in the first six month of the project, and presented in Deliverable D4.1 ("Inventory of environmental ontologies and corpora from the environmental domain").

SWEET[28] (Semantic Web for Earth and Environmental Terminology) is an ontology developed by the Jet Propulsion Laboratory of California Institute of Technology, which describe relevant knowledge in Earth Science and Environmental domains. It covers many subjects and topics, e.g. Space, Astronomy, and Biology. The latest available release is version 2.0 Beta, which is divided in 125 ontology modules, and includes more than 4500 concepts, almost 300 properties, and about 500 individuals. Among the available modules, particularly relevant for PESCaDO are the ones describing Atmosphere, Climate, Phenomena and Geology. The DL expressivity of the whole ontology is *SHOIN(D)*.

The Environment Ontology[29] is a community-based ontology for describing the environment of any organism or biological sample. It is particularly focused on the description of biome. The latest available release includes more than 1200 concepts, over 7000 individuals, but basically no properties (just 5). It is written according to the OBO file format (an alternative ontology representation language), but a derived OWL version is also available. The DL expressivity of the ontology is *ALE+*.

Another available ontology is the Weather Ontology[30], provided by the University of Aberdeen. This ontology is used by an agent to report the current weather in Aberdeen as well as a forecast a few days ahead. Albeit being smaller in size with respect to the previous two ontologies considered (around 100 concepts, 40 properties, and 40 individuals), it seems to be very focused on the application domain of PESCaDO, describing entities like *SkyCondition*, *CloudType*, *MaximumTemperature*, and *PrecipitationEvent*. It is written in DAML+OIL, but it can be translated to OWL thanks to some state of the art converters[31]. The DL expressivity of the ontology is *ALUHN(D)*.

Other relevant resources can be accessed via the Community Data Portal[32], which collects datasets, models, and ontologies (coming mainly by international institutions like the National Center for Atmospheric Research and the University Corporation for Atmospheric Research) on several topics, like Climate and Weather. Similarly, the Marine

[28] http://sweet.jpl.nasa.gov/

[29] http://www.environmentontology.org/

[30] http://www.csd.abdn.ac.uk/research/AgentCities/WeatherAgent/index.php

[31] see e.g. http://www.daml.org/2003/06/owlConversion/

[32] http://cdp.ucar.edu

Metadata Interoperability Project[33] hosts a huge repository of ontologies and vocabularies potentially relevant for PESCaDO.

If we consider just the three ontologies here presented, albeit being quite different in size and topics coverage, it is evident that in some cases they describe very related entities. Just as an example, the concept *Drizzle*, which is a kind of *Precipitation* in the SWEET ontology, is without any doubt related to the *Drizzle* concept, which is a kind of *PrecipitationEvent*, in the Weather Ontology. Ontology Alignment techniques and tools come in our help to identify and correctly represent situations like this.

7.3 State of the Art of Ontology Alignment tools and techniques

Ontology Alignment is the problem of finding correspondences between the entities of two or more ontologies. As such ontology alignment plays a crucial role in a number of knowledge and information management related tasks, including i) ontology engineering by promoting the reuse and merging of existing ontologies, ii) ontology learning by supporting the identification of appropriate positions for the introduction of new concepts/properties, iii) the integration of heterogeneous ontologies, etc. Although no specific initiative has been reported for the environmental domain, ontology alignment has been deeply investigated in the general setting, due to its criticality for many data-intensive applications, such as schema/ontology integration, catalogue matching, agent communication, web service integration, P2P information sharing, query mediation, and so forth.

Among the first initiatives to ontology alignment research fall the early Semantic Web related projects, e.g. WonderWeb[34], CROSI[35] and SWAP[36], while seminal contributions to further advances constitute the efforts undertaken within the projects Knowledge Web[37] and SEKT[38], which delivered comprehensive and insightful surveys, much as novel mapping methodologies. Recent relevant activities include TONES[39], NEON[40], BOEMIE[41], etc. Additional key resources include the book on Ontology Matching by Euzenat & Schvaiko [5], which provides a systematic analysis of the various proposed approaches aiming to serve as a uniform framework of reference, the user perspective ontology mapping tools review presented in [6], the analysis on trends and challenges for ontology matching [9], [4], [8], etc. (For a comprehensive list of pointers to relevant publications and activities, the reader is referred to the "official" ontology matching[42] site).

[33] http://marinemetadata.org/

34 http://wonderweb.semanticweb.org/

35 http://www.aktors.org/crosi/

36 http://swap.semanticweb.org/

37 http://knowledgeweb.semanticweb.org/

38 http://www.sekt-project.com/

39 http://www.tonesproject.org/

40 http://www.neon-project.org/

41 http://www.boemie.org/

42 http://ontologymatching.org/

As the extensively rich literature reveals, the field abounds of techniques and tools for the discovery of correspondences/mappings between ontologies. What is still missing though is a consensual, well-defined framework establishing the merits of the different approaches and, more urgently, a theory for which approach (combination of approaches) should be applied in which circumstances, in an effective and flexible manner.

Despite originating from an old problem (see database schema integration, federated databases, etc.) and encompassing highly extensive and active research activities, ontology alignment remains yet quite controversial when it comes to the classification of the proposed approaches, as differing viewpoints have been adopted with respect to the definition of ontology correspondences, their interpretation and discovery strategies. Generally, a correspondence can be considered a 4-uple for the form $<e_1,e_2, r,d>$, where e_1 and e_2 are two entities belonging to different ontologies, r the relation representing the correspondence holding among them, and n a confidence measure for this correspondence. Further refinements can incur, e.g. when context is taken into account and introduced in the representation.

Typical options for the relation r include *equivalence* '$=$' (two concepts represent the same entity), *subsumption* '$\subseteq$' (a concept is more specific than another one), *overlap* '$\cap$', and *disjointness* '$\perp$' We note though that a correspondence may not necessarily refer to a semantic relations such as the aforementioned, but may as well convey a generic notion of concept similarity.. The specification of such correspondences (mapping) determines the logical relations that hold between entities of different ontologies, accomplishing the alignment of the ontologies. Generally speaking, the main basic techniques involved in the process of matching (i.e. the specification of semantic overlap between two entities) can be categorised as follows.

Linguistic techniques. These techniques evaluate the similarity among ontology entities on the basis of their names and the names of their properties. They may involve lexical, syntactic, and/or semantic approaches, and encompass algorithms based on string distances, on statistical measures coming from the frequency of occurrence of a term, and/or on external resources, like WordNet.

Structural techniques. These techniques base the computation of the alignments between entities on the potential relations involving them (e.g. subsumption relation between entities). Typically, these techniques computes type comparisons or graph-based similarities.

Extensional (instance based) techniques. These techniques base the computation of the alignments between entities comparing their potential extensions. Typically, the extensions considered are the individuals belonging to concepts.

Semantic (reasoning) techniques. These techniques are based on some formal semantics (i.e. model-theoretic approaches) to justify the matching produced, e.g. using some intermediate formal ontology as a common ground on which to base the comparison.

Statistical techniques. These techniques employ statistical inference in order to calculate the probability that two entities, belonging to different ontologies, are similar or share overlapping instances. Frequency probability and Bayesian probability have been widely used to capture the notion similarity between two entities as joint probability distribution.

Each category can be further refined on the basis of the specific methodology and theory followed; the distinction however is often rather vague, since to some extend each of the aforementioned categories employs elements of the other categories. Adding to this the common practise held by the majority of existing ontology alignment methodologies and tools to combine multiple approaches, the direct classification and comparison of the existing literature becomes quite ambiguous. Moreover, since the strategy employed in order to combine the matching resulting from the various techniques and obtain a meaningful measure of correspondence/similarity, may as well be considered as an additional classification dimension.

The inclination towards the combination of different techniques is clearly justified given the complexity and challenges involved in analyzing and discovering correspondences and the level of semantics, while taking into account the varying peculiarities confronted in different application domains. Furthermore, as the different approaches entail corresponding strengths and weaknesses, they tend to constitute partial solutions rather than evolving towards a dominant one, a possibility, which given the extensive research already conducted and the challenges and issues involved, remains highly unlikely. Such observations are reflected not only by reports on individual experiences in ontology alignment research, but also by the yearly evaluation events that are organised by the Ontology Alignment Evaluation Initiative[43] (OAEI), since 2004.

Table 7.1 captures the main aspects of well-known ontology alignment tools, in terms of the types of ontological entities addressed, the employed matching strategy, and the approach followed for the combination of the individual similarity measures into a coherent, combined value. As illustrated, the combination of more than one approach is favoured, while it is worth noting that a common combination strategy involves iterative computations, where subsequent similarity measures refine the ones obtained earlier through the application of a different approach.

Tool	Ontology Elements	Matching Strategy	Similarity Combination
FOAM[44]	concept, properties, instances	linguistic, structural	weighted sum, cumulative
S-Match[45]	concepts	linguistic, semantic	iteration
PROMT[46]	concept, properties, instances	linguistic, structural	cumulative
GLUE/iMap[47]	concept, properties, instances	statistical	iteration
COMA++[48]	concept,	linguistic,	average value, iteration

[43] http://oaei.ontologymatching.org/

[44] http://www.aifb.uni-karlsruhe.de/WBS/meh/foam/

[45] http://semanticmatching.org/

[46] http://protege.stanford.edu/plugins/prompt/prompt.html

[47] http://pages.cs.wisc.edu/~anhai/projects/schema-matching.html

[48] http://dbs.uni-leipzig.de/Research/coma.html

	properties, instances	structural	
OLA[49]	concept, properties, instances	linguistic, structural	iteration

Table 7.1: Main features of selected ontology alignment tools.

The list of the afore-examined tools is by no means exhaustive, additional prominent approaches include H-Match[50], IF-MAP[51], PRIOR+[52], etc.; it indicates though illustratively the multiplicity of methodologies adopted within the existing approaches. Selecting among the existing tools and methodologies and appropriately adapting and/or extending them, depends on the specific requirements that characterise the application and domain considered each time.

Discovering ontology mappings though, constitutes only part of the challenges concerning the tasks of alignment, articulation and merging of ontologies required to support ontology reuse and engineering, semantic integration, and knowledge sharing. Consequently, besides ontology alignment algorithms, of particular interest for the purposes of PESCaDO are tools and frameworks for ontology integration and merging that allow to subsequent utilise the discovered mappings. From the aforementioned tools, FOAM enables the user to configure also ontology integration and merging tasks, while PROMT comprises in a framework for multiple ontology management including ontology merging (iPROMT), ontology versioning (PROMPTDiff), and sub-ontology factoring (PROMPTFactor), in addition to ontology mapping (AnchorPROMPT). Chimaera[53] is a system that supports the creation and maintenance of distributed ontologies on the web, addressing the merging of different ontologies and the checking activities that are required during the ontology life-cycle. Users are supported in loading knowledge bases in differing formats, reorganizing taxonomies, resolving name conflicts, browsing ontologies, and editing terms, while receiving suggestions on ontology elements that form possible candidates for merging.

MAFRA[54] allows creating semantic relations between two (source and target) ontologies, and applying such relations in translating source ontology instances into target ontology instances. FCA-Merge is a framework for ontology merging, where the tool interactively drives the domain expert in building manually the merged ontology. SAMBO[55] supports users in ontology aligning tasks by making suggestions about potential mappings, and in addition

[49] http://ola.gforge.inria.fr/

[50] http://islab.dico.unimi.it/hmatch/news.php

[51] http://www.aktors.org/technologies/ifmap/

[52] http://www.sis.pitt.edu/~mingmao/om07/index.html

[53] http://www.ksl.stanford.edu/software/chimaera/

[54] http://sourceforge.net/projects/mafra-toolkit/

[55] http://www.ida.liu.se/~iislab/projects/SAMBO/

carries out the actual merging of the ontologies and derives the logical consequences resulting from the merged operations.

Particularly relevant to PESCaDO is also the Alignment API[56] which is an API and an implementation for expressing and sharing ontology alignments. The developed alignment format is expressed in RDF, rendering it easy to futher extend it, and its motivation lies in providing a uniform representation for alignments so that they can be easily shared. EDOAL[57] (Expressive and Declarative Ontology Alignment Language), currently under revision for the implementation of the version 4.0 of the Alignment API, extends this format to enable the representation of complex correspondences, where more precise relations than that of subsumption and equivalence are required.

Relevant notions, although not directly developed within the context of ontology alignment, include C-OWL [3], an extension to OWL so as to represent bridge rules between ontologies, Distributed Description Logics (DDL) [2], which use directional mappings to preserve local terminological schemas while formalizing knowledge integration, ε-connections [7] that allow to combined DL knowledge-based, and package-based Description Logics (P-DL) [1], an approach to modular ontology language to further enhance knowledge reuse and integration.

For the purposes of PESCaDO, two main factors need to be taken into account when selecting among the available ontology alignment methodologies and frameworks. First, the peculiarities pertaining to the environmental domain and the available ontologies; in this aspect, experience from ontology aligning efforts in the medical domain may prove particularly useful, as both domains share scientific terminology much as more industrial and application oriented vocabulary. Second, given the intricate challenges encountered in ontology alignment in general, and the partial nature of the proposed approaches, appropriate strategies need to be devised so as to meet effective on one hand the project specific objectives, while on the other hand enhancing the robustness and reliability of the alignments. Critical elements underlying these goals, comprise the effective handling of imprecision, which depending on the context my appear in the form of uncertainty and/or vagueness, as well as the efficacious utilisation of the formal semantics and reasoning apparatus, especially under the realistic, scalable framework offered through distributed setting.

7.4 State of the Art of Ontology Extension tools and techniques

Ontology extension is a subtask of *ontology learning*, i.e. the acquisition of new concepts and relations between them. In PESCaDO, the problem of extending and adapting existing ontologies to the specific issues of the meteorological domain will be addressed.

In the last years, a number of Natural Language Processing techniques have been developed which try to automatically extract concepts and their relations from corpora. Among relations the most frequently targeted is the isa-relation. A widely used technique for the acquisition of this relation requires the identification of linguistic patterns expressing the relation in texts. A number of such patterns have been proposed in the literature [Hearst, 1992[58]; Hearst, 1998[59]; Mititelu, 2006]; yet, with the exclusion of (Snow et al., 2005)[60], little is reported about systematic assessments of pattern reliability as predictors of the relation. Even less is known about the applicability of such techniques to domain specific ontologies in concrete real world contexts.

[56] http://alignapi.gforge.inria.fr/

[57] http://alignapi.gforge.inria.fr/edoal.html

[58] Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545.

[59] Marti A. Hearst, 1998. *WordNet: An Electronic Lexical Database*, chapter Automated discovery of wordnet relations. MIT Press, Cambridge, MA.

[60] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, Cambridge, MA.

As for the source of information required to acquire the concepts and relations, the Web offers several advantages. For example, thanks to its huge size and heterogeneity it covers almost all existing knowledge domains. Moreover, reliable learning methods can be built based on its high degree of redundancy and the presence of publicly available search engines. Other valuable knowledge sources required for ontology extension include electronic dictionaries, linguistic databases like WordNet [Fellbaum, 1998][61] and WordNet domains (Magnini & Cavaglià, 2000)[62], as well as structured and semi-structured information. A combined method based on Google snippets and WordNet was successfully employed in the PatExpert project and, due to its high adaptability to new domains, could be tested also in PESCaDO. However, some problems may arise from the use of the Web as a knowledge source because of its highly dynamic and uncontrolled changing nature, which make it very difficult to structure information. Besides, plenty of documents are available for every possible domain, and it may be problematic to filter spam or unreliable sources.

7.5 AVAILABLE STANDARDS

Semantic Web Technologies will play a key role within PESCaDO, in particular for what concerns storage/representation of resources and ontology languages. For this reason, it is fundamental to monitor and consider the adoption of state of the art Semantic Web Technologies standards. The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, recommendations, software, and tools) to lead the Semantic Web to its full potential. The main standard Semantic Web-related technologies relevant for the PESCaDO project are:

OWL 2 Web Ontology Language (OWL 2). W3C Recommendation 27 October 2009

Resource Description Language (RDF). W3C Recommendation 10 February 2004

SPARQL Query Language for RDF. W3C Recommendation 15 January 2008.

Uniform Resource Identifier (URI): Generic Syntax - RFC 3986, STD 66 - January 2005 - Copyright © The Internet Society (2005)

7.6 REFERENCES

[1] Bao, J., *Voutsadakis*, G., *Slutzki*, G., *Honavar,* V.: Package-Based Description Logics. *Modular Ontologies 2009*, 349-371.

[2] Borgida, A., Serafini, L.: Distributed description logics: Assimilating information from peer sources. Journal of Data Semantics 1, 153–184 (2003).

[3] Bouquet, P., *Giunchiglia*, F. van Harmelen, F. , Serafini, L. Stuckenschmidt, H.: C-OWL: Contextualizing Ontologies, *International Semantic Web Conference (ISWC 2003*), 164-179.

[4] Choi N., Song I.-Y., Han H.: A Survey on Ontology Mapping Sigmod Record, 2006.

[5] Euzenat J. and Shvaiko P.: Ontology Matching, Springer-Verlag, Heidelberg, isbn 3-540-49611-4, 2007.

[6] Falconer, S.M., Noy N.F., Storey Margaret-Anne D.: Ontology Mapping - a User Survey. OM 2007.

[7] Grau, B.C., Parsia, B., Sirin, E.: Working with multiple ontologies on the semantic web. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 620–634. Springer, Heidelberg (2004).

[61] Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press.

[62] Bernardo Magnini and Gabriela Cavaglià. 2000. Integrating Subject Field Codes into WordNet. In Proceedings of LREC 2000, pages 1413–1418, Athens, Greece.

[8] Kalfoglou, Y., *Schorlemmer*, W.M. : Ontology Mapping: The State of the Art. *Semantic Interoperability and Integration 2005*.

[9] Shvaiko P., Euzenat J.: Ten Challenges for Ontology Matching, OTM Conferences (2) 2008: 1164-1182.

8 DISTILLATION OF CONTENT STRUCTURES FROM MULTILINGUAL WEB MATERIAL

8.1 DESCRIPTION OF THE PROBLEM DOMAIN

PESCaDO will develop techniques for the derivation of deep-content structures and semantic clues from environmental material. These techniques, which we call *content distillation*, will serve two purposes: 1. extraction of the functional coverage of an environmental service node based on its metadata and the content of its web pages; 2. population of environmental ontologies. In more concrete terms, such tasks will involve deciding whether a certain website is relevant to the purposes of PESCaDO and extracting specific data such as temperature, street conditions, weather forecast, etc.

The content distillation task has to be extended to all languages involved in the PESCaDO pilot use cases, namely Finnish, English and Swedish and for this reason it has to include applications and methodologies that are easily adaptable to different languages. Content will be distilled from different kinds of text, including metadata in the web pages of the environmental nodes, concrete information bulletins as offered by service-oriented nodes in the web, and considerably more abstract and linguistically complex multimodal resources such as background information repositories (for example metadata of environmental portals). Such an approach extends the existing mainstream research on automatic ontology acquisition, which focused so far only on homogeneous text corpora.

8.2 GLOSSARY/ABBREVIATION

Relation extraction	The detection and classification of semantic relationships between a set of entities identified in a text.
Frame	A prototypical conceptual structure which is evoked by specific predicates, called *lexical units*, and comprises a set of participants or semantic roles, called *frame elements*.
SRL	Semantic Role Labeling – a natural language processing task involving the detection of the semantic arguments associated with a predicate and their classification into specific roles, e.g. *Agent*, *Patient*, etc.
WSD	Word Sense Disambiguation – the task of deciding what is the meaning of a word in a specific context

8.3 DESCRIPTION OF TECHNOLOGIES AND TOOLS THAT ARE TYPICALLY USED FOR CONTENT DISTILLATION

The vast majority of the state of the art systems for ontology population through content distillation use term identification and/or named-entity recognition and classification in order to locate instances of concepts (and in some cases also instances of relations between concepts), which are then integrated into the given ontology. Some systems populate an ontology with instances of both concepts and relations (as, e.g., Artequakt [Artequakt], WEB->KB [WEB-

>KB], SOBA [Buitelaar et al., 2006]), while others focus only on relation instances extraction (such as Adaptiva [Adaptiva], LEILA [Suchanek et al., 2006]). KnowItAll [KnowItAll] processes only concept instances, while other systems (as, e.g., Artequakt and SOBA) do not learn terms and synonyms, but use publicly available NLP tools for this task. Some systems (e.g KnowItAll) include a term/synonym extraction engine, but require extraction patterns to be provided by the user. WEB->KB, Adaptiva, and LEILA include an adaptable term/synonym extraction engine, which can be taught either with the help of concept/relation lexicalisations or through concept/relation instance examples. For term/synonym extraction, machine learning seems to be the technique adopted by the majority of the systems. All these systems rely mostly upon linguistic information and therefore either use an external, publicly available term/synonym extraction engine or require manually constructed patterns as input. LEILA, in addition, uses filtering based on statistical approaches. The machine learning based systems either use statistical methods to identify terms or perform automated pattern extraction. Similar to the state of the art methodologies for ontology learning, the state of the art for content distillery (or ontology population) methodologies are developed for general discourse.

In order to understand the peculiarities of environmental information and to extract the domain-specific terminology, the collection and analysis of available multilingual environmental corpora, glossaries and thesauri is needed. A variety of portals and sites, as well as terminological resources are available for the domain in different languages. For example, the European Environment Information and Observation Network [EIONET] has developed the online GEMET thesaurus [GEMET], that includes a list of themes, such as climate, air and transport, grouping different concepts that are provided with a definition and the translation in 27 languages. More than 6,000 descriptors have been encoded so far, which can be freely downloaded in different formats. On the website of the European Environment Agency[63], instead, the Environmental Terminology and Discovery Service (ETDS)[64] allows users to look for environmental terms, browse their definition and some associated sample images, as well as their translation in various languages. Another useful resource is the Environmental Dictionary[65] (EnDic), developed by the Finnish Meteorological Dictionary, that includes 90,000 search words in nine languages and covers environmental terminology about nature conservation, wastewater treatment, environmental policy and others. EnDic is available together with MetDic, containing 9,000 meteorological entries in Finnish, Swedish and English.

In order to process such resources, a set of standard tools for NLP is needed, in particular a morphological analyzer and a syntactic parser. While for English several resources are freely available, such as the Stanford parser [Klein and Manning, 2003], Dan Bikel's statistical parser [Bikel, 2000], and the TextPro suite [Pianta et al, 2008] among others, such tools are less widespread for Finnish and Swedish and have poorer performances if compared to English. The MaltParser [Nivre et al., 2006] is a data-driven dependency parser provided with pretrained models for English and Swedish. The model for Finnish could be obtained by training the parser with a Finnish dependency Treebank. Another interesting application is the Helsinki Finite-State Transducer [HFST] developed at the Department of General Linguistics of the University of Helsinki. It includes an open source implementation of lexica and NLP tools for English, Swedish, Finnish and French. Other taggers and parsers for English, Swedish, Finnish and other European languages have been developed by the Connexor [Connexor] company but they are not freely available.

Given that text resources in the environmental domain are already available, two kinds of content distillation are considered by the state of the art: on the one hand, a *shallow content distillation* for general content analysis by extraction of the most important keyword-concepts, and on the other hand a *deep content distillation* that can provide a more accurate semantic analysis for explicit representation of background knowledge.

63 www.eea.europa.eu/

64 http://glossary.eea.europa.eu/

65 http://mot.kielikone.fi/mot/endic/netmot.exe?UI=ened&height=165

As for shallow content distillation, a widely used technique is *keyphrase extraction* from text, i.e. the identification of the most important concepts in a document. This allows for the identification of the main topics dealt with in the document and for a quick classification of the content (for different applications and methodologies see for example [Frank et al., 1999; Lawrie at el., 2001; and Medelyan & Witten 2006]).

All keyphrase extraction approaches are based on the same basic steps: 1) select the candidate phrases 2) decide whether candidates are keyphrases and 3) possibly post-process the keyphrase list.

Two well-known state-of-the-art tools for keyword extraction exploit supervised machine learning techniques: GenEx [Turney, 2000] and KEA[66]. GenEx utilizes a set of parameterized heuristic rules that are tuned to the training corpus by a genetic algorithm, while KEA (and KEA++) uses a naïve-Bayes learning method to induce a probabilistic model from a training corpus. An on-line keyphrase extraction service is offered also by a number of websites, for example Extractor[67], Keyword Analysis Tool[68] and Dublin Core metadata editor[69]. Another tool for shallow content distillation is the *Kx* tool, integrated in the *TextPro* natural language processing suite[70] developed by FBK. This tool would be particularly suitable for content distillation in the PESCaDO project because it is language independent and can be easily applied to Finnish or Swedish texts, given a training corpus in that language. Besides, it is possible to tailor the keyword extraction task to the environmental domain by boosting as keyword candidates the domain-related collocations.

On the other hand, deep content distillation deals mostly with a *relation extraction* task, i.e. the detection and classification of semantic relationships between a set of entities identified in a text. A wide variety of relation classification schemes exist in the literature, reflecting the needs and granularities of various applications. Rosario ad Hearst [2001], for example, developed a system for relation extraction focused on the medical domain, introducing 13 domain-specific classes. Stephens at al. (2001) proposed 17 very specific classes for classifying relations between genes. Some researchers only investigate relations between named entities or between noun head and modifier [Rosario and Hearst, 2001; Nastase and Szpakowicz, 2003], while others have a more general focus. For example, Moldovan et al. [2004] use a 35-class scheme to classify relations in various phrases. A novel approach was successfully proposed and experimented in the PatExpert[71] project, i.e. the set of relations were defined based on a small set of pre-defined prototypical situations or *frames* [Fillmore, 1976] and the entities considered corresponded to the semantic roles or *frame elements* involved in every frame.

As shown by the results of the SemEval-2007 task 04 “Classification of Semantic Relations between Nominals” [Girju et al. 2007], the usage of WordNet and Google queries constitute the two most widely-used approaches to validate a systems’ output in this task. In addition, most systems exploit also syntactic information (see for example [van Hage & Katrenko, 2007, Giuliano et al., 2007]) and SVM-based machine-learning techniques [Butnariu and Veale, 2007; Bedmar at al., 2007; Aramaki et al., 2007].

In PESCaDO, a set of relations should be selected in order to capture the relevant information required by the

66 http://www.nzdl.org/Kea/

67 http://www.extractor.com/

68 http://seokeywordanalysis.com/

69 http://www.ukoln.ac.uk/metadata/dcdot/

70 http://textpro.fbk.eu/

71 PATExpert project, *http://www.patexpert.org/*, IST ***FP6*** 028116

environmental service nodes. Since the use of *frame* information in the deep content distillation step of the PatExpert project has proved to deliver promising results when applied to the patent domain for a limited set of relations, it may be worth investigating a domain-specific adaptation of such analysis to the environmental domain of PESCaDO.

8.4 AVAILABLE STANDARDS AND EVALUATION INITIATIVES

The most important research initiatives trying to organize and evaluate the objectives of content distillery (or extraction) are the SemEval evaluation campaigns, which focus on the task of undestanding the meaning of a word in context, and the TAC (Text Analysis Conference)[72] track about knowledge base population (formerly Automatic Content Distillery initiative (ACE)), which deals with methodologies of detection and normalization of specific types of information according to predefined conceptual models.

SemEval is organized by Senseval[73], an international organization devoted to the evaluation of word sense disambiguation (WSD) systems and other semantic analysis. The last campaign launched by Senseval in 2010 includes 17 tasks, some of which are of potential interest for PESCaDO: cross-lingual word sense disambiguation, automatic keyphrase extraction, multi-way classification of semantic relations between nominals, word sense induction, all-words word sense disambiguation in the environmental domain. The task list shows that word sense disambiguation (and its variations) is seen as a relevant issue for language understanding, and that multilingual systems deserve increasing attention from the Natural Language Processing (NLP) community.

The standard resource for WSD is WordNet [Fellbaum, 1998], a lexical database of English, in which nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct concept. Synsets are linked by means of conceptual-semantic and lexical relations. A related resource is WordNet Domains [Magnini & Cavaglià, 2000], which labels each WordNet synset with a domain label taken from a pre-defined domain hierarchy. Since WordNet is not available for Finnish, some research work in PESCaDO will be devoted to the investigation of strategies to automatically derive it from existing resources (either in English or in Finnish).

The TAC (Text Analysis Conference) is organized by the American National Institute of Standards and Technology (NIST) and consists of a set of tracks about different NLP tasks. One of the tasks is called “knowledge base population” and its main objective is to develop automatic content extraction technologies to support automatic processing of human language in text form. The initiative was formerly called ACE (from 2003 to 2007) and was devoted to three types of source texts, namely newswire, broadcast news and newspaper. Since 2009, the track has been divided into two subtasks, namely entity resolution given a knowledge base and relation extraction. The main goal is to detect specific content objects mentioned in a text, recognize some selected information about these objects and merge them into a unified representation for each detected object. Relevant content objects include: entities, values, temporal expressions, relations and events.

In PESCaDO, such tasks would be particularly interesting, for example the ability to recognize generic entities (e.g. spatial locations, meteorological phenomena), relations between them (e.g. pollens cause allergies) and complex events (e.g. rainy weather makes the streets slippery for cyclists). Also named-entity recognition referring to locations and temporal expressions related for example to weather forecasts would be relevant tasks within PESCaDO.

8.5 GAPS, MISSING FUNCTIONALITIES AND OPEN ISSUES IN THE STATE OF THE ART

One of the open issues with regard to PESCaDO is the multilingual extension of existing techniques and resources to Finnish and Swedish. In fact, most of the language-dependent tools required for content distillation are

72 http://www.nist.gov/tac/tracks/index.html

73 http://www.senseval.org/

already available for English, however they may need some adaptation to the environmental domain. On the contrary, few NLP tools for Finnish and Swedish are freely available. First they should be tested in order to see if their performance complies with the standard achieved for English. If not, machine learning-based tools should be trained for the two languages, after collecting a sufficient amount of annotated data. Besides, while English and Swedish are Germanic languages and present a certain degree of typological similarity, Finnish is part of the Finno-Ugric group of languages, which means that it has a very complex inflectional structure and a completely different grammar compared to Germanic languages. The morphological component of NLP tools used for content distillation will be carefully tuned to take into account the peculiarities of Finnish.

Another issue of the content distillery part is the analysis of complex multimodal resources. So far, the SemEval evaluation campaigns and the standards provided in the ACE/TAC program have not taken into account multimodal material, where text is combined with graphics, tables and background information repositories provided by web portals. Also mainstream research on automatic ontology acquisition has focused so far on homogeneous text corpora. This means that important information can be extracted also from different sources and not just by processing body text of html pages. This fact is an open issue and would require a new multimodal approach, in which table extraction and image processing techniques could complement traditional natural language processing.

8.6 REFERENCES

[Adaptiva] http://www.aktors.org/technologies/adaptiva/

[Aramaki et al., 2007] E. Aramaki, T. Imai; K. Miyo, and K. Ohe. 2007. UTH: SVM-based Semantic Relation Classification using Physical Sizes. In *Proceedings of SemEval Task 4*.

[Artequakt] http://www.artequakt.ecs.soton.ac.uk/

[Bedmar et al., 2007] I. Segura Bedmar, D. Samy, and J. L. Martinez. 2007. UCM3: Classification of Semantic Relations between Nominals using Sequential Minimal Optimization. In *Proceedings of SemEval Task 4*.

[Bikel, 2000] D. Bikel 2000. A statistical model for parsing and word-sense disambiguation. In *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and very large corpora*, Hong Kong.

[Buitelaar et al., 2006] Buitelaar, P., P. Cimiano, A. Frank, S. Racioppa. 2006. SOBA: SmartWeb Based Ontology Annotation. In Proceedings of the International Semantic Web Conference.

[Butnariu and Veale, 2007] C. Butnariu and T. Veale. 2007. UCD-S1: A Hybrid model for detecting semantic relations between noun pairs in text. In *Proceedings of SemEval Task 4*.

[Connexor] http://www.connexor.eu/

[EIONET] www.eionet.europa.eu/

[Fellbaum, 1998] C. Fellbaum. 1998. WordNet. An electronic Lexical Database. MIT Press.

[Fillmore, 1976] Fillmore, C. J. 1976. Frame Semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language*, pages 20–32. Blackwell Publishing.

[Frank et al., 1999] E. Frank and G.W. Paynter and I. Witten and C. Gutwin and C.G. Nevill-Manning, 1999. Domain Specific Keyphrase Extraction, Proceedings of the 16th International Joint Conference on AI, pp.668--673.

[GEMET] http://www.eionet.europa.eu/gemet

[Girju et al., 2007] R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th Semantic Evaluation Workshop (SemEval-2007)*.

[Giuliano et al., 2007] C. Giuliano, A. Lavelli, D. Pighin and L. Romano. 2007. FBK-IRST: Kernel Methods for Semantic Relation Extraction. In *Proceedings of SemEval Task 4.*

[HFST] http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/index.shtml

[Klein and Manning, 2003] D. Klein and C. D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems* 15 (NIPS 2002). MIT Press. pp. 3-10

[KnowItAll] http://www.cs.washington.edu/research/knowitall/

[Lawrie at el., 2001] Dawn Lawrie and W. Bruce Croft and Arnold Rosenberg, 2001. Finding Topic Words for Hierarchical Summarization, Proceedings of SIGIR 2001, New Orleans, Louisiana, USA.

[Magnini and Cavaglià, 2000] B. Magnini and G. Cavaglià. 2000. Integrating Subject Field Codes into WordNet.In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.) *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, pp. 1413-1418.

[Medelyan and Witten 2006] O. Medelyan, I. H. Witten, 2006. Thesaurus based automatic keyphrase indexing. In: Proc. of the JCDL 2006, Chapel Hill, NC, USA.

[Moldovan et al., 2004] D. Moldovan, A. Badulescu, M. Tatu, D. Antohe and R. Girju. 2004. Models for the semantic classification of noun phrases. In *HLT/NAACL 2004 Workshop on Coputational Lexical Semantics*, pp. 60-67.

[Nastase and Szpakowicz, 2003] V. Nastase and S. Szpakowicz. 2003. Exploring noun-modifier semantic relations. In Fifth International Workhop on Computational Semantics (IWCS), pp. 285-301.

[Nivre et al., 2006] Nivre, J., J. Hall and J. Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, May 24-26, 2006, Genoa, Italy, pp. 2216-2219.

[Pianta et al., 2008] E. Pianta, C. Girardi, R. Zanoli. 2008. The TextPro tool suite. In *Proceedings of LREC.* Marrakech, Morocco.

[Rosario and Hearst, 2001] B. Rosario and M. Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of EMNLP*. 2001, pp. 82-90.

[Suchanek et al., 2006] Suchanek, F., G. Ifrim, and G. Weikum. 2006. LEILA: Learning to Extract Information by Linguistic Analysis. In Proceedings of the Second Workshop on Ontology Learning and Population: Bridging the gap between text and knowledge.

[Turney, 2000] P. Turney, 2000. Learning Algorithms for Keyphrase extraction. *Information Retrieval*, vol. 2, pp. 303-336.

[van Hage and Katrenko, 2007] W. Robert van Hage and S. Katrenko. 2007. UVAVU: WordNet Similarity and Lexical Patterns for Semantic Relation Classification. In *Proceedings of SemEval Task 4*.

[WEB -> KB] http://www.cs.cmu.edu/~webkb/

9 USER-ORIENTED REASONING AND DECISION SUPPORT STRATEGIES

One of the aims of PESCaDO is to provide to the end-user support in taking decisions related to environmental aspects. Typically, support to take a decision comes by effect of an interactive process, in which the following macro-activities take place:

1. the user describes his/her planned activity or problem and formulates a request to obtain decision support;
2. the decision support system (DSS) reacts to the request of the user and, by means of some reasoning service, it tries to fulfil it;
3. the DSS communicates to the user the answer to his/her request.

This can be reasonably considered the scenario also for PESCaDO. In particular, each of these activities hides some key challenges that PESCaDO has to adequately address:

As described in Section 7, the PESCaDO system will rely on ontological knowledge to provide user-decision support. Hence, for the system, a request is typically described in terms of ontology axioms; conversely, the users usually formulate requests in natural language or in a graphical way. Therefore an intermediate representation of the user request/problem is needed. For this reasons, one of the challenge of PESCaDO is to define an appropriate and expressive problem description language (PDL), tailored for the environmental domain, which allows the communication of the request from the user to the system.

The PESCaDO system has to implement semantic reasoning services to adequately fulfill the user requests. The reasoning services will be based on a combination of logical reasoning services, content filtering services, and the data made available by the environmental services. The implementation of the reasoning services has to take into account two important aspects which are characteristics of the environmental domain: the system will have to work with a large quantity of data, and the provenance of these data, i.e. their accuracy, may vary.

The PESCaDO system, combining ontology reasoning techniques, content selection techniques, and automatic text generation techniques will have to identify the content considered relevant to provide (and communicate) adequate decision support to the user.

Below, we briefly describe if and how similar problems have been investigated in the literature. In particular focusing on describing some available environmental DSSs and existing proposal of PDLs. To the best of our knowledge, neither environmental specific user-oriented DSSs nor environmental specific PDLs exist. We also describe some state-of-the-art reasoning systems and approaches to work with large quantities of data, time and uncertainty.

9.1 GLOSSARY/ABBREVIATION

PDL	Problem Description Language
DSS	Decision Support System
DL	Description Logics
KB	Knowledge Base

9.2 OVERVIEW OF STATE-OF-THE-ART DECISION SUPPORT SYSTEMS

The state of the art abounds of environmental (or environmental related) decision support systems. It has to be noticed however, that they are more institution/community-oriented decision support systems, rather than user-oriented DSS as in the PESCaDO vision. Furthermore, differently from PESCaDO, most of them rely only on numerical models rather than on semantic web technologies.

As a first example of environmental DSS, we include the one described in[74]. In this work, an optimisation-based environmental decision support system (EDSS) was developed for supporting sustainable rural development. The system key components are a dynamic database system, a graphical user interface, and a mixed integer linear programming model. The developed EDSS has been applied in the use case of the Yongxin County, located in Jiangxi Province, in China.

The ECOSIM project[75] was focused on building a model-based information and decision support system for urban environmental management. The system integrates data acquisition and monitoring systems, GIS, and dynamic simulation models in a flexible client-server architecture based on standard protocols. The system relies on several numerical models, including atmospheric wind-field model, air pollution dispersion and air chemistry model, ground and surface water level and pollution model, coastal water pollution model, and traffic and traffic emission model. One of the main components of the platform, is an embedded rule-based expert system (which uses near-natural language rules), which assists users in the definition of decision and input variables, as well as in the interpretation of results.

OntoWEDSS[76] is a decision support system for wastewater management. In addition to rule-based reasoning and case-based reasoning, OntoWEDSS uses an internal knowledge-base and inference mechanisms to process information about a wastewater treatment plant. The goal of the system is to support a human manager's decisions in maintaining the correct operation of the wastewater treatment plant. The output of the system are statements about actions to be taken by the manager, based on a diagnosis/prediction of the ongoing/future state of the treatment plant. At the core of OntoWEDSS lies the WaWO ontology, a hierarchically-structured set of terms and relations describing the domain of wastewater treatment.

An on-going project[77] is the development of the Semantic DSS for Palm Oil Industry, run by the RANN Consulting, Malaysia. The aim of the Semantic DSS for Palm Oil Industry is to support plant managers in making crucial and timely decisions, taking into consideration factors ranging from costing, environmental issues and policies governing the operations of the palm oil industry. To provide adequate reasoning and to support effective decision making, the system integrates usage of Semantic Technology through ontologies, context-management and web-services, augmented with Bayesian Network (BN) and Fuzzy Logic to handle the uncertainty of data.

---

[74] An optimisation-based environmental decision support system for sustainable development in a rural area in China - G. H. Huang ab; X. S. Qin c; W. Sun c; X. H. Nie c; Y. P. Li d

[75] http://www.ess.co.at/ECOSIM/

[76] Ontowedss - an ontology-based environmental decision-support system for the management of wastewater treatment plants - Luigi Ceccaroni (2001, PhD Thesis) - url: http://www.tesisenxarxa.net/TESIS_UPC/AVAILABLE/TDX-0225103-180427//thesis.pdf

[77] Semantic Decision Support System (DSS) and Portal for Palm Oil Industry - Sagaya Sabestinal

Semantic Technology Conference 2009 - San Jose, CA - June (2009) url: http://www.semanticuniverse.com/articles-semantic-decision-support-system-dss-and-portal-palm-oil-industry.html

DIAGNOZA_MEDIU[78] is a rule-based expert system that provides qualitative information to a DSS used in an environmental protection management domain. This expert system is based on a knowledge base (which comprises data coming from environmental and meteorological database, and also forecasting data), an inference engine, an explanation module, a knowledge acquisition module, and a user interface. The system allows to handle uncertain knowledge, in the sense that a set of terms (corresponding to linguistic certainty values), is used to express the user's degree of confidence in the facts stored in the knowledge base. The system has been already applied for air pollution analysis and control in some urban region.

9.3 OVERVIEW OF STATE-OF-THE-ART PROBLEM DESCRIPTION LANGUAGE

Problem Description Languages have been widely used in several implemented system to foster both human-system communication and communication between different components within a system. As an example of the latter, see the Brick Problem Description Language[79]. In this brief survey, we will focus on PDL for human-system interaction.

In[80], the problem description language used in PROUST is described. PROUST is a program diagnosis system that aids novice programmers to check and fix non-syntactic bugs in the code they implemented. A PDL is used in PROUST in order to render the requirements written in the natural language problem description in a form processable by the system. In particular, in PROUST the problem description is written in a simple notation (containing a few special keywords), and consists of the name of the problem, together with objects and goals definitions.

PDLs are also used in network surveillance expert systems, like shown in[81]. In this case, the PDL adopted follows an English-like syntax, which resembles vaguely the one of a programming language. The problem descriptions considered in this system involve elements like rules, actions, events, and conditions. The users modify or create problem descriptions with the help of a customization tool, and the resulting descriptions are saved in ASCII files called problem scripts. These files are then compiled into the record structures used by the expert system.

In[82], a PDL is used in the context of a system implementing a development tool for a rule-base language called ORBS. In this system, a limited domain-specific problem description language has been implemented with the requirement of taking into account automatic design and code generation from problem description. The PDL was based on NIKL, a member of the family of KL-One knowledge representation languages. However, in this system, the user does not write problem descriptions directly in NIKL, but via a graphical editor which hides the syntax of the language to him/her.

[78] A case study of knowledge modelling in an air pollution control decision support system - Mihaela Oprea - AI Communications (2005) Vol.18 (No.4)

[79] EvoCAD: EVOLUTION-ASSISTED DESIGN - Pablo Funes, Louis Lapat and Jordan B. Pollack (2000) url: http://www.demo.cs.brandeis.edu/pr/buildable/evocad/aid00/

[80] Intention-Based Diagnosis of Novice Programming Errors - Lewis Johnson (1986) Morgan Kaufman Publisher, ISBN 978-0934613194

[81] Customization of network surveillance expert systems - Andrzej Bieszczad , Tony White (1992) http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=276592&userType=inst

[82] Development tools for rule-based systems - Stephen Fickas (in "Expert systems: the user interface" (1987) - http://portal.acm.org/citation.cfm?id=40529)

A more recent approach is the one presented in the MONET EU project[83]. The aim of MONET is to demonstrate the applicability of the latest ideas of the Semantic Web to the world of mathematical software, and in particular mathematical services. The Mathematical Problem Description Language (MPDL) is an XML-based application which allows a user to describe a mathematical problem in terms of its inputs and outputs, and pre/post-conditions. The descriptions created are then grouped into a library for future reuse. See MONET Deliverable D14[84] for more details.

9.4 Overview of state-of-the-art reasoning systems handling large data, time and uncertainty

Intelligent decision support for the environmental domain requires reasoning over environmental knowledge in ontologies and knowledge bases. In PESCADO, the reasoning will be based on a combination of ontological reasoning, content filtering, and structured data made available by the environmental services which constitute the background knowledge base. Although no reasoning system specifically tailored to knowledge provided by environmental service nodes exists in the state of the art, there are several mature and application-domain independent formalisms and reasoning tools available, many of them also used in real world applications to solve heterogeneous problems (e.g. information retrieval, decision support).

Firstly, we quickly report some state-of-the-art scalable RDF repositories. Boca[85] is a scalable RDF repository developed by the IBM Adtech group in Cambridge. Boca's key component is the server capable of storing millions of RDF triples (DB2 database). Boca RDF triples can be easily accessed client-side via the Jena APIs[86] and the RDF SPARQL query language.

Sesame[87] is an extensible and configurable open source Java framework for storage and querying of RDF data. The basic version directly supports SPARQL and SeRQL querying, and RDF Schema inferencing. Sesame can also be interfaced with other tools via APIs or the RESTful HTTP protocol.

OWLIM[88] Semantic Repository a high-performance semantic repository developed in Java. It basically adds a storage and inference layer on top of the Sesame RDF framework. Its semantic is user configurable, and the default version supports unconstrained RDF Schema inferencing plus most of OWL Lite. It is available mainly in two versions: a free one (SwiftOWLIM - LGPL license - scalable to about 10 million triples), and a commercial one (BigOWLIM - scale to billions of triples).

Modularized knowledge base can help in improving the efficiency of working (mainly querying) with large knowledge bases. In a modularized knowledge base, the data is organized in modules, and a query can be run against the relevant modules only. The Contextualize Knowledge Repository (CKR)[89] is a recent contribution in this direction.

[83] IST-2001-34145, see MONET homepage for more details: http://monet.nag.co.uk/monet/

[84] http://monet.nag.co.uk/monet/publicdocs/monet-msdl-final.pdf

[85] http://ibm-slrp.sourceforge.net/2006/11/20/boca-the-rdf-repository-component-of-the-ibm-semantic-layered-research-platform/

[86] http://jena.sourceforge.net/

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

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

[89] Context Shifting for Effective Search over Large Knowledge Bases – Mathew Joseph, Luciano Serafini and Andrei Tamilin - In Proceedings of the Workshop on Context, Information And Ontologies (CIAO-2009) Heraklion, 2009

In the CKR, the RDF triples are partitioned into parts (*contexts*), in each of which knowledge is conforming with a certain selected state of the world, e.g., is about a certain topic, period of time, geographical place, etc. Hence, each context can be characterized according to some structured dimensions (time, location, and topic in the current implementation), the value sets of which can be organized in a hierarchy. Search services can be run on a specific context only, or in relevant related contexts via a mechanism called *context shifting*. The current implementation of the CKR is based on Sesame RDF store.

Supporting efficient reasoning on large quantity of data is the goal of several state-of-the-art approaches. Among them, the Large Knowledge Collider (LarKC), an FP7 EU project[90] that aims at providing support for Web-scale reasoning by developing a platform for massive distributed incomplete reasoning. One of the results of the project is the development of MaRVIN[91] (Massive RDF Versatile Inference Network), a parallel and distributed platform for performing RDF(S) inference on a network of machines based on a peer-to-peer model. MaRVIN can be scaled to arbitrary size (the scalability of the system is achieved by adding computing nodes to the network) and it is not tailored to a single specific reasoner.

Other state-of-the-art scalable reasoning systems are DRAGO[92], which addresses the problem of reasoning with multiple distributed ontologies, pairwise interrelated by semantic mappings, and Owlgres[93], an open source, scalable reasoner for OWL2 (limited to DL-Lite expressiveness) which combines Description Logic reasoning with the data management and performance properties of an RDBMS.

In the last few years, several works have investigated the extensions of Description Logics (DL) and OWL language to cope with time and uncertainty (see e.g.[94]). We report here some of the main results and achievements in these directions since, due to the nature of the environmental domain, they are relevant for PESCaDO.

Concerning time, the work on MT–ALCO and OWL-MeT[95] has extended DL and OWL to work with topological and metric properties of time. In particular, MT–ALCO is an extension of *ALCO* Description Logics which integrates the hybrid metric temporal logic *MT* (which includes temporal operators like *future n*, *somepast*, *allfuture*, etc.). OWL-MeT is the variant of OWL language formally grounded in MT–ALCO. A reasoner (called Pellet-MeT) for OWL-MeT is also available: as suggested by its name, the implementation is based on Pellet[96] OWL reasoner.

Concerning imprecision, several attempts have been taken to address the issue of modelling vague knowledge in the semantic web. An example is Fuzzy OWL[97] (or f-OWL) that extends the OWL web ontology language with fuzzy set theory (a mathematical framework for covering vagueness), in order to be able to capture, represent and reason

[90] http://www.larkc.eu/

[91] http://www.larkc.eu/marvin/

[92] http://sra.fbk.eu/projects/drago/index.html

[93] http://pellet.owldl.com/owlgres/

[94] Reasoning within Fuzzy Description Logics - Umberto Straccia - Journal of Artificial Intelligence Research, Vol. 14: 137-166, 2001.

[95] http://ermolayev.com/owl-met/

[96] http://clarkparsia.com/pellet

[97] Fuzzy OWL: Uncertainty and the Semantic Web - Giorgos Stoilos, Giorgos Stamou, Vassilis Tzouvaras, Jeff Z. Pan and Ian Horrocks - In Proc. of the International workshop on OWL: Experience and Directions (OWL-ED2005)

with information that may be imprecise or vague. With respect to standard OWL, f-OWL allows to assign a membership degree (a value from 0 to 1) in the definition of facts of the knowledge base.

A different approach that, addresses imprecision from a probabilistic uncertainty perspective is the one considered in BayesOWL[98]. Following a probabilistic approach, BayesOWL extends OWL using Bayesian networks (BN – a widely used graphic model for probabilistic interdependency) as the underlying reasoning mechanism and probabilistic model. In the BayesOWL approach probabilistic constraints are represented as OWL statements, the OWL ontology is then translated into a BN directed acyclic graph, and finally available probability constraints are incorporated into the conditional probability tables (CPTs) of the translated BN. Ontology reasoning is then performed within the translated BN as Bayesian inferences.

A third approach, still founded in Bayesian Networks, is the one proposed by PR-OWL[99]. PR-OWL is a probabilistic extension to OWL that provides a framework for authoring probabilistic ontologies. By definition, a probabilistic ontology is an explicit, formal knowledge representation that comprehensively describes knowledge about a domain, and the uncertainty associated with that knowledge. PR-OWL is a Bayesian ontology language based on Multi-Entity Bayesian Networks (MEBN) logic that provides the means to express first-order probabilistic theories.

Umberto Straccia has worked extensively on Fuzzy extensions to Description Logics. Besides theoretical contributions, he has developed two software applications[100] that may be relevant for PESCaDO project: the fuzzyDL System, which is a DL Reasoner supporting Fuzzy Logic and fuzzy Rough Set reasoning, and the fuzzyRDF System, which supports fuzzy reasoning over RDF triples to which an uncertainty degree is attached. Other related literature implementations, include FiRE[101], which extends the DL *SHIN* with fuzzy set theory semantics, and DeLorean[102], which supports the translation from a fuzzy rough ontology language into a classical ontology language, allowing subsequently the invocation of a classical DL reasoner.

Concerning RDF knowledge bases, some other approaches that deal with uncertain RDF triples are worth mentioning. Fuzzy RDF[103] (and Fuzzy RDF Schema) proposes a syntactic and semantic extension of RDF. In Fuzzy RDF, a statement is a couple (value, triple), where triple is a traditional RDF triple, and value is a real number in the interval [0,1]. A fuzzy RDF reasoner, based on Sesame, was also implemented.

Analogously, URDF[104] (Uncertain RDF) extends RDF with the capability to express uncertainty by allowing to associate RDF formulas with probabilities. Actually, as claimed by the author, it also supports the semantics of RDFS

[98] Z. Ding, Y. Peng and R. Pan, BayesOWL: Uncertainty modeling in Semantic Web ontologies. In: Z. Ma, Editor, Soft Computing in Ontologies and Semantic Web, Studies in Fuzziness and Soft Computing vol. 204, Springer (2006).

[99] PR-OWL: A Framework for Probabilistic Ontologies - Paulo Costa, Kathryn B. Laskey - Proceedings of the Fourth International Conference on Formal Ontology in Information Systems, November 2006.

[100] http://gaia.isti.cnr.it/~straccia/software/

[101] http://www.image.ece.ntua.gr/~nsimou/FiRE/

[102] http://webdiis.unizar.es/~fbobillo/delorean.php

[103] Mauro Mazzieri. A fuzzy rdf semantics to represent trust metadata. In Proceedings of the 1st Italian Semantic Web Workshop: Semantic Web Applications and Perspectives (SWAP 2004), 2004.

[104] Dealing with uncertainty in the semantic web - Rienstra, T.D. (2009) - Master's thesis, Univ. of Twente.

and part of OWL. The extension works on top of the Semantic Web layer. In URDF, reasoning is performed by combining rule-based inference on RDFS/OWL knowledge, with Bayesian inference.

Another tool available is Incerto[105], a probabilistic reasoner for the Semantic Web. The system is based on Markov logic, and implements ideas coming from the statistical relational learning domain. The tool supports not only reasoning about uncertainty in the Semantic Web, but also how to learn that uncertainty from available resources (e.g. ontology individuals, textual corpus, and web search engines). The proposed tool (and approach) has some limitations arising from the application of Markov logics that could limit its application in various domains: inefficient performances when dealing with transitive or cardinality restrictions, and the impossibility to define uncertainty information about facts (individual class/property membership).

A further tool for probabilistic reasoning is Pronto[106], an extension of Pellet that offers core OWL reasoning services for knowledge bases containing uncertain knowledge. In addition to the standard Pellet features, Pronto allows to add probabilistic statements to OWL ontologies, to infer new probabilistic statements from probabilistic ontologies, and to explain the results of probabilistic reasoning. Pronto is available open source.

Finally, we conclude this section by mentioning Alchemy[107], a tool providing a bunch of algorithms for statistical relational learning and probabilistic logic inference. Alchemy is based on the Markov logic representation, a powerful language that combines first-order logic and probabilistic graphical models by attaching weights to first-order formulas and treating them as templates for features of Markov random fields. Alchemy is available Open-source, and a set of API is available to extend and integrate the tool in external applications.

10 USER-SYSTEM INTERACTION TECHNOLOGIES

10.1 GLOSSARY

HCI	Human Computer Interaction
VA	Visual Analytics
InfoVis	Information Visualization
PDL	Problem Description Language
Dialog	Part of a GUI that requests input from the users. Commonly realized as a window with input elements, e.g. containing a form that can be filled and accepted or declined.
GUI	Graphical User Interface
PATExpert	EU founded project IST-028116 researching the possibility to enhance patent search through linguistic, semantic, and visual means

[105] http://code.google.com/p/incerto/

[106] http://pellet.owldl.com/pronto

[107] http://alchemy.cs.washington.edu/

DSS	Decision Support System
MCV	Multiple Coordinated Views – Different views on the same object of investigation, that are coordinated in the sense of e.g., exchanging selection events to cross-highlight selected sections.
Treemap	A space-filling visualization technique for hierarchical data
Focus and Context	A class of interaction techniques that focus a region within the view while keeping the surrounding context visible for interpretation, usually in a reduced and/or skewed representation.
Star Plot	Also called Radar Chart, Spider Chart, or Kiviat Diagram. A circular line diagram where every dimension is a line leaving the center and the values are depicted as a line strip that crosses all legs.
Dynamic Query Sliders	A widget for letting users set the constraint value for result filters, which automatically updates the result list instantly while being used.
Widget	Portmanteau word for a "Window Gadget". An interactive element of a GUI.

10.2 Description of the Problem Domain

Combining semantic reasoning with user oriented decision support in the domain of environmental services poses a variety of challenges. The formal logic of semantic reasoning has to handle the imprecise nature of user requests (potentially expressed in natural language) together with the uncertainty of environmental forecast data. Adding to this complexity is the large amount of available environmental data as well as the possibility to create unforeseeable combinations of environmental aspects and task in the user's requests.

Within PESCaDO, human computer interaction (HCI) techniques will be developed to address this complexity and bridge the gap between the world of informal user problems and formal logic. PESCaDO will focus on HCI in three specific areas: (i) Visual interfaces for problem description languages (PDLs) in decision support environments; (ii) visual support in query expansion techniques; (iii) the use of Visual Analytics (VA) techniques for service confidence/uncertainty metrics determination

10.3 Visual Interfaces for Problem Description Languages

The formulation of environmental decision support queries faces three fundamental problems:

1. The query language needs to be as precise as possible to be able to calculate an answer in an appropriate period of time.
2. Yet, the query language needs to be flexible enough to allow for formulating many different kinds of requests.
3. The users need to be able to estimate if a query at hand resembles their intended request for the decision support system.

To support users in expressing their information need, all three topics should be addressed. Purely textual and formal input of queries in the problem description language is unbearable for new users because they lack the

knowledge about the PDL. Allowing natural language input, however, is not an option due to its imprecise nature. Both problems can be faced with dialogs containing a set form because it shows the available options and only allows correct input. But this solution comes at the cost of the desired flexibility which is one of the main reasons for utilizing semantic reasoning within PESCaDO.

As our results from PATExpert (IST-028116) showed, hybrid approaches seem promising for this task. Combining controlled textual input with an interactive display of the current query allows the users to better understand meaning of the stated request while the request at hand is formal and precise. This enhanced understanding allows users to adjust and expand the queries efficiently and therefore improving flexibility.

Visual support for querying languages has been developed in many communities, especially for querying relational databases. A comprehensive overview of traditional visual support approaches – from a simple query structuring to icon based environments that allow drag and drop creation of query statements – is presented in the survey by Catarci et al.[108]. Traditional approaches are especially tailored to the query language or the of data repository they are supposed to support – e.g., querying XML documents[109]. Some recent works are of a higher relevance to PESCaDO. Dongilli et al. (2004)[110] provide in the context of SEWASIE (IST-34825) an interface for the formulation of precise queries on heterogeneous data in the economy domain. Morris et al. (2004)[111] present a visual querying language for spatial databases, which works on a low abstraction level not suitable for intuitive problem description. In PATExpert, an approach has been developed by USTUTT to build queries in a visual interactive manner for a combination of semantic, image-oriented, similarity-based, and plain text based patent search; cf. Figure 10.1.

[108] Tiziana Catarci, Maria Francesca Costabile, Stefano Levialdi, Carlo Batini. "*Visual Query Systems for Databases: A Survey*", J. Vis. Lang. Comput. 8(2): 215-260, 1997

[109] Braga, D.; Campi, A., "*A graphical environment to query XML data with XQuery*", Web Information Systems Engineering, 2003, pp. 31-40, 2003

[110] Paolo Dongilli, Enrico Franconi, Sergio Tessaris: "*Semantics Driven Support for Query Formulation*", Description Logics, 2004

[111] Andrew J. Morris, Alia I. Abdelmoty, Baher A. El-Geresy, Christopher B. Jones: "*A Filter Flow Visual Querying Language and Interface for Spatial Databases*", GeoInformatica 8(2):107-141 (2004)

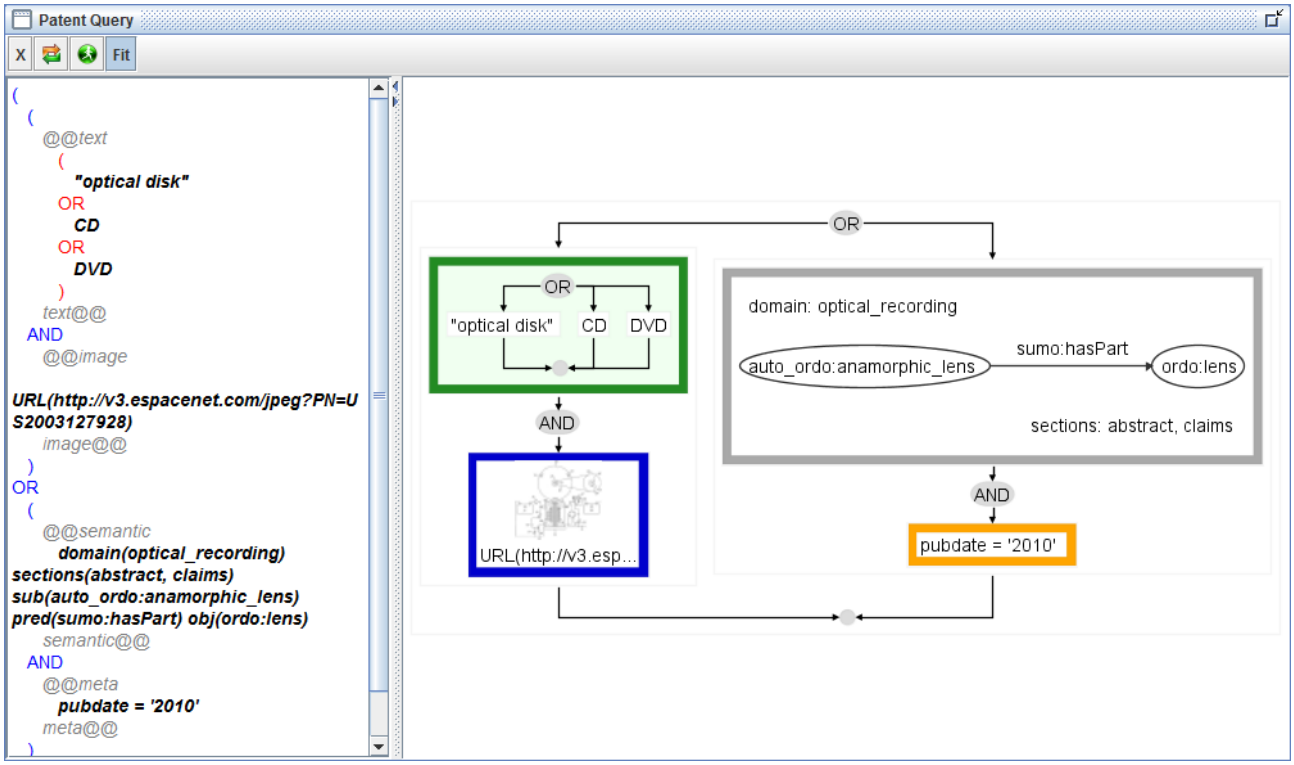


Figure 10.1: Query formulation in PATExpert

10.4 VISUAL SUPPORT FOR QUERY EXPANSION

Decision support systems (DSS) feature a rich interaction between the users and the system and usually exceed single question and single answer scenarios. This can be illustrated by the following examples.

The answer proposed by the system changes the original plan of the user and therefore necessitates an updated decision support query.

The outcome of the query does not relate to the original problem of the user which also necessitates a change of the query.

The user is not pleased with the proposed rational of the decision support and seeks further information about alternatives or the underlying data on which the decision-making is based.

For this interaction process the system has to provide a representation of the answer of the system as well as a linkage to the related query to be able to adjust and resubmit it efficiently. The representation of the decision support result strongly depends on the available data within the answer and therefore the aspects involved for solving the query. Here, the visualization module will rely mainly on the output of PESCaDO's multilingual user tailored environmental information synthesis and delivery module.

For the linkage to the original query, the system needs to be able to understand the aspects of the result of the system in order to intelligently support the user to incorporate these aspects at the right location of the problem description. Adding self expressiveness to the result and its representation at a system level is not only helpful for creating means to feeding back aspects to the query but also for allowing interaction on the result itself in order to increased comprehensibility.

10.5 VISUAL ANALYTICS FOR CONFIDENCE METRIC DETERMINATION

Visual Analytics[112] (VA) is a fledgling scientific discipline that combines the fields of search, machine learning, data mining, and mathematical modeling with the fields of information visualization and HCI. Most of the VA systems developed so far are tailored to a special application domain. In most cases these systems are dealing with emergency response[113], social networks[114], finance transaction[115], and situation awareness[116].

These domains are not selected by chance. VA is most useful in scenarios where large quantities of heterogeneous data are at hand and the aspect that is searched for is not clearly defined or distributed across multiple data sources. This necessitates a very flexible approach that utilizes the capabilities of the human perception and cognition system to tweak semiautomatic algorithms that can scan through vast amounts of data.

The domain of environmental services also features many different data types from many data providers. These data need to be linked and processed in order to

- create the data base for offering a value added service on top of existing services,
- estimate the confidence of single services or derived services, or
- manage, assess and control the quality of offered services.

Especially the confidence metric determination and the orchestration of services show the same characteristics as the domains mentioned above. This makes them promising targets for a VA driven approach. Here, semiautomatic methods can calculate the confidence values for a currently selected set of environmental services and metric parameters to give instant feedback which helps the user to judge the suitability of his choice for the task at hand.

VA approaches that are more generic are Polaris[117] and DataMeadow[118]. Polaris is a system to query, analyze and visualize multidimensional databases. It is an example of successful VA software that is actually used in the field. Part of its success lays in its familiar pivot table based interface as well as the ubiquity of relational databases in business life. Polaris allows the user to easily correlate, aggregate, and group different data types and deliver a

---

[112] J. J. Thomas and K. A. Cook, **"***Illuminating the Path: The Research and Development Agenda for Visual Analytics***",** National Visualization and Analytics Ctr, 2005. [Online]

[113] Andrienko, G.; Andrienko, N.; Bartling, U., "Visual Analytics Approach to User-Controlled Evacuation Scheduling", Visual Analytics Science and Technology, 2007, pp.43-50

[114] Bilgic, M.; Licamele, L.; Getoor, L.; Shneiderman, B., "D-Dupe: An Interactive Tool for Entity Resolution in Social Networks", Visual Analytics Science And Technology, 2006.

[115] Schreck, T., Tekušová, T., Kohlhammer, J., and Fellner, "Trajectory-based visual analysis of large financial time series data", SIGKDD Explor. Newsl. 9, 2007

[116] Aragon, C.R.; Poon, S.S.; Aldering, G.S.; Thomas, R.C.; Quimby, R., "Using visual analytics to maintain situation awareness in astrophysics", Visual Analytics Science and Technology, 2008, pp.27-34

[117] Stolte, C.; Tang, D. & Hanrahan, P. "Polaris: a system for query, analysis, and visualization of multidimensional relational databases", IEEE Transactions on Visualization and Computer Graphics, 2002, 8, 52-65

[118] Elmqvist, N.; Stasko, J. & Tsigas, P. "*DataMeadow: A Visual Canvas for Analysis of Large-Scale Multivariate Data*". In: Information Visualization, 2008, Volume 7, 18-33

compact visual representation of the result while the whole configuration can be rapidly and incrementally changed to come to the desired query and output.

DataMeadow is a system that focuses on the tight integration of query and result representation in a shared canvas. The canvas contains the history of the analysis session and the outcome of the query as a set of so called DataRoses and bar charts or pie charts. The DataRoses display the data themselves as star plots but also act as interactive and dynamic query sliders to filter the data as well as set operators to combine data sources.

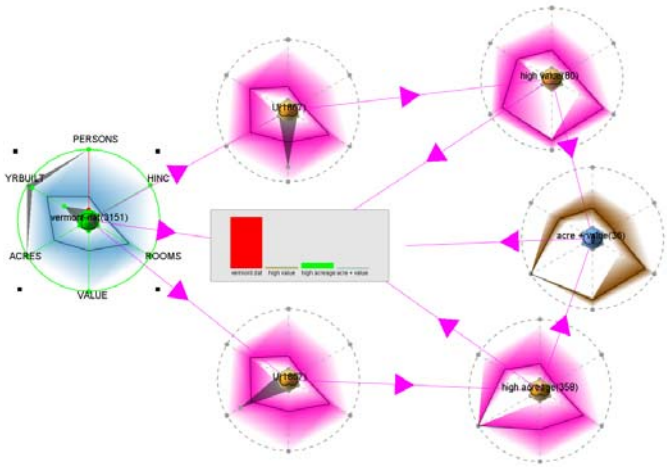


Figure 10.2: DataMeadows example from Elmqvist et al.118

10.6 Information Visualization Reference Model

Card et al.[119] as well as Chi[120] proposed a reference model for displaying information. Here, a short introduction to the reference model by Card et al. Is given to exemplify the standard procedure in information visualization.

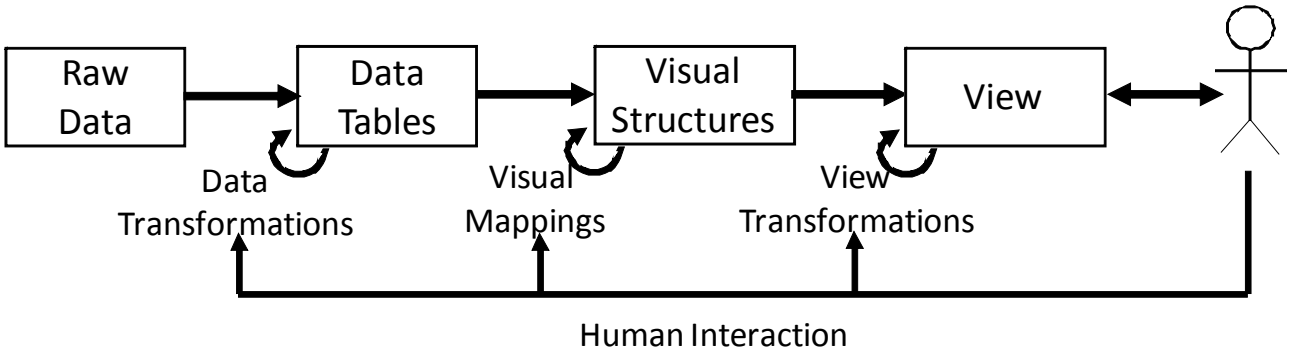


Figure 10.3: Reference model for visualization according to Card et al.119

As a first step, raw data in various forms needs to be accessed and transformed it into one or more data tables. In the tabular form every entity of the data domain is a row in a table and has multiple attributes stored in its columns. This standard representation allows reuse of algorithms and is typically easier to map to visual structures

[119] Card, S. K.; Mackinlay, J. D. & Shneiderman, B. "*Readings in information visualization: using vision to think*", Morgan Kaufmann Publishers Inc., 1999

[120] Chi, E.H. "*A taxonomy of visualization techniques using the data state reference model*". In: IEEE Symp. Information Visualization (InfoVis 2000), pp. 69 – 75, 2000

than the raw data itself. Common transformation steps during the Data Transformation stage are filtering, aggregating, or calculating new values. For example, implicit hierarchies have to be transformed to parent-child relations stored in tabular form in this step.

The next step maps the data tables to visual structures. Most often this is an extension of the data tables to include visual attributes like size, shape, color, location, orientation, etc. These attributes determine the actual visualization type, be it a bar chart, treemap, node-link diagram, or alike. Here, expressiveness is of special importance. A visual structure is expressive if all and only the data of the data tables are expressed in the visual structure.

Finally the visual structures are made available to the user through a view. Thinking of the visual structures as a base, view transformations can augment the visual variables to explore the data. For example, a viewport can be changed (e.g. scrolling over a document), the visibility of entities can be modified (e.g. slicing through a volume), or location mappings can be distorted to create lens like focus and context views.

Of utmost importance for creating a visualization application for Visual Analytics is the possibility for the user to interact with all aforementioned transformation steps. While being able to change view transformation properties is nowadays available in many applications, it most often is not the case for defining the visual mappings or even the data transformations.

11 MULTILINGUAL USER-TAILORED ENVIRONMENTAL INFORMATION SYNTHESIS AND DELIVERY[121]

The synthesis (or "automatic generation") of environmental information, ideally tailored to the needs of the addressee, is an application of what in Natural Language Processing is known as Report Generation (RG). RG is a domain-specific application of Natural Language Generation (NLG). Since the early days of NLG, report generators have been developed for a number of different domains – among them labour market [Rösner, 1986; Iordanskaja, et al., 1992], stock market [Kukich, 1983; Reiter and Dale, 1990], weather [Goldberg et al., 1994; Coch, 1998; Yao, 1998; Sripada et al., 2003], air quality [Busemann and Horacek, 1997; Wanner et al., 2007], patient histories [Bontcheva and Wilks, 2004] and team game commentaries [Robin, 1994]. A number of them passed beyond the prototypical implementation, some reached an operational state; cf., e.g., FoG [Goldberg, et al., 1994], MultiMeteo [Coch, 1998], AutoText [Bohnet et al., 2001], and Narrative Engine [Harris, 2008].

11.1 CHARACTERISTIC FEATURES OF REPORT GENERATION

RG reveals the following characteristics:

(i) its input is composed of structured data: numerical data time series or an abstract content representation;
(ii) it implies assessment and interpretation of the input data, which can involve summarization in accordance with some (possibly implicit) criteria of relevance;
(iii) it takes the restrictions of the sublanguage of the field in question into account.

Report generators are typical "data-to-text" and "concept-to-text" applications. They take as input structured data (e.g., one or several numerical time series measured by external devices or an abstract content representation), which require, at least to a certain extent, assessment and interpretation. Guided by criteria of relevance to established norms or to the addressee, assessment and interpretation often involve summarization of the

[121] Section 11 is based on [Wanner, 2010]

data/content. Strictly speaking, data assessment and interpretation is not necessarily an RG task: it is needed for any content presentation mode – be it a text, a table, a curve, or a diagram. In AI, this task has often been attributed to expert systems [Giarratano and G. Riley, 2005]. Therefore, in many advanced report generators, the data assessment and interpretation module is not considered part of the generator. However, some authors also argue for the inclusion of assessment and interpretation into the RG system architecture [Reiter, 2007]. While assessment converts data into content or makes inferences in a knowledge base, it does not include selection of the content that is to be communicated in a report in order to satisfy the needs of the user. Content selection constitutes a separate task, which can be omitted if the content of the entire input structure is to be verbalized, as in the case of SEMTEX [Rösner, 1986], LFS [Iordanskaja et al., 1992], and FoG [Goldberg et al., 1994]. In a few report generators, user profile or context criteria are taken into account to drive the selection; cf., for instance, MARQUIS and SumTime-Mousam. In TEMSIS [Busemann and Horacek, 1997], the user selects the content directly via an interactive interface. From the discussion in the previous subsection of the idiosyncrasies of sublanguages with which RG has to deal at all levels of the linguistic description, it is obvious that report generators need to take these idiosyncrasies into account in order to produce a naturally sounding text.

As far as the architecture of report generators is concerned, it is not different from the architecture common in general discourse NLG. Most often, this is a pipeline architecture [Reiter and Dale, 2000]. The number of modules and the distribution of tasks among the modules varies (which is not different from general discourse NLG generators). However, as already indicated above, the nature of the tasks addressed by report generators is nearly always the same: (1) data assessment and interpretation, (2) content selection, (3) discourse planning, and (4) linguistic realization. The use of multiple modi, i.e., tables and graphics along with the texts, has been largely neglected so far in RG.

In the following subsections, we discuss what RG starts from and how the tasks (1) – (4) tend to be addressed in RG.

11.2 What do the report generators start from?

RG uses a number of data and knowledge sources and implies certain preprocessing stages, notably some data preprocessing and data interpretation.

11.2.1 Data and Knowledge Sources

In general, RG draws upon three different data and knowledge sources: input data, background knowledge of the domain and user models. However, not all generators use all three sources. Especially user models are often absent in simpler report generators.

11.2.1.1 Input data

As pointed out above, the input to RG are structured data. The data format and the level of abstraction may be rather different. Thus, LFS [Iordanskaja et al., 1992], MARQUIS, TEMSIS, SumTime-Turbine [Yu et al., 2007], etc. take numerical time series; MIAKT [Bontcheva and Wilks, 2004] starts from a structure related to an ontology; PLANDOC [McKeown, et al., 1995], BT-45 [Portet et al., 2009], etc. draw upon both numerical time series and ontologies. In contrast to general discourse NLG, RG hardly ever starts from linguistic semantic or syntactic structures. Numerical time series are very common as input to report generators, which clearly lies in the nature of the task: numerical series monitored over time call for an interpretation, assessment of relevance and a summarized verbal presentation.

The size of the series, i.e., the amount of data to be preprocessed by the assessment and interpretation shell, may vary significantly. Thus, turbine monitoring series are much larger than, e.g., air pollutant series. However, the complexity of the assessment is not necessarily proportional to the size of the series. Thus, a major share of the assessment of any numerical series is a mathematical curve analysis in that it contains, e.g., (i) determination of the start and end values within the considered interval of the series, (ii) detection of the significant relative and absolute

minima and maxima; (iii) identification of significant positive and negative gradients; (iv) evaluation of the difference between the maxima/minima and the predefined thresholds within the considered interval, etc.

Conceptual structures instantiated from ontologies or inspired by ontology representation formalisms are other "natural" input structures in RG. Some report generators use ontologies, either already as initial input, as MIAKT or for intermediate structures, as BT-45.

11.2.1.2 BACKGROUND DOMAIN KNOWLEDGE

In addition to the dynamic input, most of the report generators draw upon some static background knowledge, which is either represented explicitly in terms of a knowledge base, lists or tables or incorporated into the processing procedures. This knowledge may concern domain-specific content interpretation, static information to be included when certain contextual conditions are met and language conventions that are to be observed. A typical example of static background information are legal notices in the air quality bulletins as produced by TEMSIS and MARQUIS: specific threshold pollutant concentrations are assigned legal notices, which must be included without modification into the report when these concentrations are reached. Cf. such a legal notice for the ozone threshold of 180 μgr/m^3:

> *Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit prolonged or heavy exertion outdoors.*

The language conventions concern all levels of RG; they can be domain-oriented or cultural. Thus, domain communication knowledge can be instrumental for the realization of the discourse structure [Rambow, 1990] and also influence the syntactic structures of the sentences in the report. However, most illustrative is the domain and culture dependency of the vocabulary. For instance, in the weather domain, the notion of `hot' depends on the cultural perception (coined by the geographical location); in the air quality domain, the perception of air quality in general and of pollutant concentrations in particular varies from one region to another – which is even reflected in the environmental regulations of each country that associate air quality index scales with labels such as "good", "bad", "satisfactory", etc. The interpretation and naming of the time intervals of a day is also cultural. For example, the Spanish *mañana* 'morning' extends more or less until 2pm and the *tarde* 'afternoon' until 8 or 9pm. The German *Morgen* 'morning'' can go until 12:00, and the *Nachmittag* `afternoon' until 5pm at the latest. The diverging interpretation is occasionally rejected by the vocabulary; for instance, in German, a special term for `time before noon' – *Vormittag* – is available, while in English, French, etc. this is still morning.

11.2.1.3 USER MODELS

Advanced RG is increasingly applied to domains that require a variation of the content and the language style in which this content is presented depending on the user model. In general, a user model in NLG can be considered as being composed of the profile of the user and the history of generation. The profile captures the user's characteristic features that are relevant to generation; the history contains the information that has already been communicated to the user (either in the current or in the previous sessions). As a rule, report generators do not protocol the history.

The profile of a user may have the following dimensions: (i) expertise, (ii) need for specific information, (iii) cultural background, and (iv) interpersonal content bias, i.e., personal content interpretation preferences. The expertise dimension is the most obvious and the oldest dimension in NLG; cf., e.g., [Paris, 1993; Zukerman and Litman, 2001]. The standard range is 'expert – intermediate – novice' or a more detailed variant thereof. In RG, instead of a clear range, other (domain-specific) typologies that are tailored to the different user types may be more appropriate. For instance, the MARQUIS user typology distinguishes in this respect between air quality experts, health professionals, public administrations and average citizens. The expertise dimension is, as a rule, used to guide content selection on the detail scale. Thus, a novice is contented with an overview while an expert requires details. On the other hand, an expert knows that between the concentration of ozone and the concentration of nitrogen dioxide a reciprocal correlation holds (and thus does not need an explicit mention of it), while an average citizen does not. The expertise dimension may also constrain lexicalization and influence the discourse structure. For instance, for an

expert, the relation between the statement on the concentration of a pollutant and the statement that the air quality reached a health threatening level is appropriately realized as IMPLICATURE, while for an average citizen as ELABORATION.

The dimension that specifies the need for specific information is supposed to capture the distinct content aspects that are of relevance to different users. In MARQUIS, a jogger, an individual with weak heart conditions, and an individual with a respiratory disease will receive an air quality bulletin of the same degree of detail, but with deviating content. This is because ozone is especially critical to individuals with weak heart conditions (even if its concentration is only moderately elevated), fine dust particles (PM_{10}) are of particular relevance to patients with asthma, and a jogger must be made aware of high concentrations of ozone, carbon dioxide and PM10. In SumTime-Mousam, the need of specific weather forecast information is correlated with the location of the oil rig on which the user is located.

The dimension of cultural background may be rather important in report generators that address users from different cultural regions. For instance, in Spain, users are less sensitive to the topic of air quality than in Finland and users in Finland have a different perception of low temperatures than users in Spain. Such cultural idiosyncrasies must be considered in RG during content selection, discourse planning, and lexicalization. The dimension of the interpersonal content bias is a dimension that can be dispensed with in “objective” reporting on, for instance, gas turbine monitoring, weather, stock market exchange, etc. However, it turned out to be essential for RG in domains that are per se subjective – as, e.g., game commentaries. Thus, experiments demonstrated that a soccer game commentary is judged by an addressee to be of higher quality if the content nuances and language are adapted to his personal preferences for a team – which confirms the work by Hovy in early narrative NLG [Hovy, 1988]. Traditionally, any user is assigned a predefined profile. However, when the diversity of information that can be communicated is rather high and/or the user profiles cannot be unambiguously assigned information preferences, the possibility of a personification of the profile by the users is important. For instance, in MARQUIS, a new user is first assigned a default profile, which can be personalized during the registration procedure.

11.2.2 DATA ASSESSMENT AND INTERPRETATION

RG involves a data preprocessing stage that is occasionally considered part of text planning [Kittredge and Polguère, 2000]. However, it is very different in nature from the other tasks related to text planning and it could be even considered as being outside RG. The first, basic, task in this stage that may be needed is the transformation of the input data format into a more convenient format or the relation of the input data to data already available to the generator. For instance, quantitative time series tables may be mapped onto equivalent XML-structures (as, e.g., in AutoText [Bohnet et al., 2001] or MARQUIS). Strictly speaking, this task does not involve any assessment or interpretation.

The second task is the reinterpretation of the input data in terms of a different scale or with respect to a predefined reference date. For instance, in the weather domain, the numerical value of the speed of the wind in km/h may be mapped onto a qualitative scale (`light', `moderate', `strong', …) or onto the Beaufort scale – depending on the user. Similarly, in the air quality domain, the actual concentration of a pollutant substance (measured in μgr/m3) must be projected onto the corresponding index scale (ranging, for instance, from 1 to 6) and onto a qualitative scale (‘low’, ‘moderate’, …). Furthermore, a given concentration, temperature, labour market figure, etc. is often compared to either a prominent figure detected in the past (as, e.g., the number of unemployed in SEMTEX [Rösner, 1986] and LFS [Iordanskaja et al., 1992]) or to a predefined reference (as, e.g., a legally fixed threshold concentration in TEMSIS and MARQUIS) and the result of the comparison is incorporated into the content.

The third task is the identification of patterns within the input data and assignment of meaning to these patterns. Thus, ANA [Kukich, 1983] derives from a half hourly price time-series of a stock ‘decrease’ and ‘increase’ patterns. Similarly, in LFS the increase or decrease of the employment rate is determined via the evaluation of the numerical change between two consecutive months. MARQUIS’s assessment module performs a mathematical curve

analysis on the pollutant concentration time-series in order to identify the decreases/increases and the gradients thereof, local maxima and minima, etc. and identifies semantic relations (such as `cause', `part-of', `sequence', etc.) between conceptual configurations – similar to BT-45, which also detects 'cause', 'includes' and 'associates' relations. In SumTime-Mousam, the input data are segmented in that linear intervals in the data are identified before the segmented data are mapped onto a conceptual representation. In the Streak generator [McKeown et al., 1995; Robin, 1994], no patterns are identified, but quantitative data are assigned conceptual fact structures that explicitly represent their meaning (e.g., that a given numeric score is a win or a loss).

The fourth task is the abstraction of patterns – as, e.g., in ANA, where from the elementary 'increase' and 'decrease' patterns of a number of stocks, such complex conceptual configurations as 'broadly-based decline in the market' are derived. SumTime-Turbine contains several pattern abstraction algorithms along the temporal dimension.

The fifth task, which may be interrelated with the previous three, is the assessment of the relevance of the derived content to the users. In MARQUIS, the task consists in distilling all content that may be relevant to any of the users registered to the system. In BT-45, explicit importance ratings are assigned to the derived content.

11.3 TEXT PLANNING FOR REPORT GENERATION

Text planning in NLG traditionally involves content selection and discourse planning. Content selection deals, as the name of the task suggests, with the identification of the content that will be communicated in the text to be generated. Discourse planning defines the discourse structure of the text under construction and decides on the (linear) order in which the discourse elements are to be presented. Both tasks can be carried out simultaneously by the same mechanism or separately. This is not different in RG – although in the past, it has often been assumed that the report genre has a stereotype discourse structure which is best captured in terms of text schemas [39] that are rigid to an extent that reduces the task of text planning to the minimum. This is about to change. Depending on the concrete report generator, both tasks are dealt with in separate (sub)modules or together in one single module; as a matter of fact, in most RG-systems, it is one module. The result of text planning is a text plan, which is usually already packaged sentence-wise, although its degree of detail depends on how elaborated the linguistic realization module is (see below).

11.3.1 CONTENT SELECTION

Content selection in PESCaDO is to a major extent dealt with in the scope of user-oriented decision support provision; cf. Section 9. However, an important aspect of content selection is outside the scope of reasoning: discourse-driven content selection, i.e., content selection that follows discourse criteria to include or to omit content. For instance, the empirical study of the Finnish corpus of air quality bulletins may reveal (and, in fact, it does) that the overview of the air quality in a region is as a rule complemented by a justification (as, e.g., *Air quality is poor this morning in Helsinki's street canyons because the wind is weak and thus does not dissipate the nitrogen dioxide originating from the morning traffic*), that it is usually not made explicit what kind of particles are in the air, only "particles" are mentioned, etc. This type of content selection is in the literature dealt with in the context of natural language text (or, more precisely, report) generation [Reiter and Dale, 2000].

As mentioned above, content selection may be realized simultaneously with discourse structuring or apart. Two main planning strategies have been implemented to do content selection and discourse structuring at the same time: schema-based planning and the dynamic discourse structure planning.

Schema-based planning as suggested by McKeown [1985] presupposes the availability one or several predefined (usually by domain experts) discourse structure patterns that predetermine the type of information presented and the order in which it is presented. For instance, we can assume that in meteorological bulletins, the default is to present first the state of the skies, then the precipitation, then the temperatures, then (if applicable) visibility, wind, and so on. An explicit realization of such a pattern is followed, for instance, in the webpage of the Catalan

meteorological service: *http://www.meteocat.com/mediamb_xemec/servmet/marcs/marc_prediccio.html*. The order can vary from country to country and some additional content can be included or some can be omitted; cf., e.g., a different order, with the mention of humidity, dew point, pressure, etc. in a shorter bulletin by the Finnish meteorological service *http://www.fmi.fi/saa/paikalli.html?place=Helsinki*. Several patterns can be defined and used depending on the state of affairs (e.g., in the case of extreme weather conditions, these conditions are always presented first), and the patterns can foresee some optional or alternative content elements, but the principle remains the same: it is predefined what type of content is to be displayed and in which order this is to be done, such that the content selection task is reduced to the retrieval of this content from the data or knowledge base.

Kittredge et al. [1991] suggested a more dynamic version of the discourse schemata which allow for a flexible selection of subschemata.

Dynamic discourse structure planning is usually based on a specific discourse structure theory – for instance, the Rhetorical Structure Theory, RST [Mann and Thompson, 1988] – which assume that the discourse structure of a text is, formally, a connected tree (or, more generally, connected graph) of discourse elements, such that between the governor node and the dependent node a discourse relation holds. Fragments of the graph can be summarized to one "hypernode", such that the discourse relation may hold between an individual discourse element and a fragment of the text. The discourse relations (such as CAUSE, ELABORATION, JUSTIFICATION, etc. in the case of RST) stem from a predefined restricted set that has been derived from empirical studies. The dynamic planning strategies assume that between the content units from which the content for presentation is to be selected discourse relations have already been introduced.

The first formalization and implementation of the dynamic discourse planning using RST has been proposed by Hovy [1991,1993]. Each RST-relation is expressed as a bipartite (nucleus-satellite) plan operator. The planning starts from a designated content element CE_{Init} in the KB. In the first iteration, plan operators are applied to content element pairs (CE_{Init}, $CE_{Init\rightarrow}$), where $CE_{Init\rightarrow}$ is a content element connected to CE_{Init} by a discourse relation. $CE_{Init\rightarrow}$ in pairs that fulfil operator-specific application conditions are selected for communication and, thus, a fragment of the discourse structure is created. In the subsequent iterations, elements in the created discourse structure that are, according to predefined criteria, "growth points" of the structure are taken as initial elements such that the operators can be applied to the pairs they enter into.

Marcu [1997] proposes bottom-up RST-relation driven discourse planning strategies. Starting from elementary semantic statements (units or facts), Marcu presupposes that rhetorical relations between the statements are already predefined. Exploiting the canonical ordering between the elements of these relations to ensure local coherence and adjacency constraints between the semantic statements, he first builds up a coherent linear sequence of the semantic statements that are to be verbalized and derives then from that sequence a discourse tree.

Nearly all state of the art work on RST-based text planning uses an intentional plan operator mechanism that constructs an explicit intentional and discourse structures. Cf. the seminal work by [Moore & Paris, 1993] on planning of advisory dialogues and some follow up proposals, such as, e.g., [André et al., 1993]. In the discourse on numeric time series in general and on air quality time series in particular, several general purpose intentions can be identified. Thus, the general intention of the speaker is to let the addressee (or "hearer" in the terminology of intentional planning literature) know about the content selected as relevant to him – which can be encoded using [Moore & Paris, 1993]'s notation as (KNOW ?h ?proposition). The intention of the speaker must arguably also depend on the hearer's level of knowledge. For instance, to general public, the dependency of AQI on the concentrations of primary pollutant substances can be assumed as unfamiliar. Therefore, the speaker's intention must be to make the hearer comprehend the relation: (COMPREHEND ?h REL(?e1 ?e2)), (BELIEVE ?h REL(?e1 ?e2)), or (PERSUADE ?h REL(?e1 ?e2)). The same applies to the influence of meteorological conditions on the concentration of the individual pollutants (and thus also the AQI). A professional, who can be assumed to know the interrelation between AQI, pollutant concentrations and meteorological and other contextual conditions, does not need to be made believe it.

The speaker's intention must thus be to simply let him know the content of each schema element: (KNOW ?h ?e1), (KNOW ?h ?e2).

A number of works on text planning separate the tasks of content selection and discourse structuring. For instance, O'Donnell et al., [2001] proposes a bottom up strategy similar to [Marcu, 1997] in which content selection is driven by the notion of relevance. After content selection, statements are mapped onto text nodes (or, information spans, in RST-terminology), between which RST-relations are introduced; only one RST-relation is considered for a pair of text nodes. A series of local trees is built up, with each local tree covering a subset of statements. The local trees are linked together to form one global coherent discourse tree.

Duboue and McKeown [2003] adopt a statistical approach based on a corpus of biographical summaries automatically paired with semantic data contained in a knowledge base. They focus on descriptive texts which realize a single, purely informative goal. Texts are linked to semantic data using simple anchor-based procedures, where the values for certain attributes of the input data are matched in the target text in order to establish links. Their aim is to develop a system that can automatically acquire constraints for the content selection by analyzing variations in the data and how these variations influence changes in the text. First, the semantic data is split into clusters and language models based on bi-grams are derived for the texts associated to each cluster. The language models are then used together with flattened version of the input frames to generate a training set for a supervised rule learner categorization tool and thus obtain the content selection rules.

Barzilay and Lapata [2005] also propose a statistical approach but in a different domain (American football). Content selection is addressed as an optimization problem using a graph representation of the content that accounts for all contextual dependencies between data. Dependencies are extracted from an automatically aligned corpus of content in a relational database and text by applying machine-learning classification and statistical inference. The content selection algorithm decides on the selection of database entities (i.e. rows in a table), which constitute the basic data units of their approach. Similar to Duboue and McKeown, the aligned corpus is obtained through anchor-based matching using attributes of databse entities whose values are deemed to constitute reliable anchors for alignment, namely numbers and proper names. A standard supervised classifier based on boosting (BoosTexter) is applied to each database entity (row) in isolation, using its attibute values (columns) as features, and obtaining predictions on the selection of the entities together with weights whose magnitude express the confidence of the classifier in the prediction. The next step is to capture contextual constraints in the selection of information. An initial pool of pairwise links between database entities is built by linking entries that share one or more attribute values. A filtering step is applied to the initial pool to remove spurious links identified through a statistical analysis on the distribution of predictions. The resulting set of links indicate contextual constraints in the selection of pairs of entries, be it that both units need to be selected together or, on the contrary, that each can be selected separately but not both simultaneously. By assigning weights to each link, a weighted undirected graph is built and a min-cut optimization algorithm is applied to obtain a single optimal solution to the selection of a set of database entitites.

Kelly et al. [2009] attempt to improve upon Barzilay and Lapata's methods by offering a refined algorithm for data-to-text alignment, selecting input data with a finer granularity, and showing that holding back certain data from the machine learner and reintroducing it later on can improve results. Texts (reports of cricket matches) and data (scorecards) are drawn from on-line sources. The algorithm for alignment exploits the idiosincracies of the domain and consists in the execution of consecutive steps that attempt to establish links the system is most certain of first, accumulating evidence for other more uncertain links. Where Barzilay and Lapata approach decided on the selection of whole rows of tables from a relational database, Kelly et al. aim to select individual row/column cell references, a more difficult task. The authors also have to deal with input data that is not completely structured; the scorecards extracted for cricket matches consist of a collection of unconnected tables containing statistics for a match, rather than a fully structured relational database. To enforce a stronger structure, attributes found in the extracted tables are categorized into a set of eight cross-table categories, each category emulating a database table. For the machine learning part, instead of selecting rows first and then finding constraints, the same BoosTexter classifier is trained

with whole scorecards with selection labels for each cell rather than whole rows. The resulting classifier is not only able to output predictions on individual attribute values but also accounts for the whole context of the scorecard, that is, performs collective content selection. Finally, the authors explore improvements to the training of the classifier where some information is held back (e.g. player names) during a first training round and reintroduced later for a second training round.

Evaluation has taken a prominent role in statistical approaches to content determination, where researchers can take advantage of corpus of text paired with data to perform quantitative evaluation. Barzilay and Lapata introduce a majority baseline in which whole data tables are selected if they are verbalized (aligned to text) more than half of the time. Kelly et al. extend this by considering different levels of granularity in their evaluation: whole table, rows/groups of a table, and individual cells. They also introduce a baseline whereby indicidual attribute values (cells) are selected if they occur with high frequency across the training data. Nevertheless, it can be argued that qualitative evaluation by domain experts typically performed in experts systems is still advisable for content selection tasks [Mellish and Dale, 1998].

11.3.2 Discourse Planning

The primary task of discourse planning in RG has been so far discourse structuring, i.e., determination of the order between the paragraphs as well as between the messages within the individual paragraphs. This is because it is observed that in many domains, the report discourse structure does not vary significantly and can thus be captured by a predefined *schema*. Cf. a basic schema for the AQ domain as used in MARQUIS:

1. AQ-index and rating
2. Primary pollutants
3. Secondary pollutants
4. For each primary pollutant
    4.1 concentration or index
    4.2 rating
    4.3 VIPs and VICs
    4.4 alert (if applicable)
    4.5 health risks
5. Archive information

5.1 pollutant concentrations/indices over D days

6. Forecast for each pollutant selected by the user
    6.1 expected concentration / index
    6.2 justification
    6.3 alert (if applicable)

The schema means that first, AQ-index and the rating of air quality is reported on; second, information on primary pollutant substances is given; etc.

FoG, ANA, AutoText and a number of other generators implement a rigid schema-oriented planning such that the order of the paragraphs in a report they produce and the order of the statements within each paragraph are predetermined. However, in applications where content selection is dynamic to a large extent, this strategy can be problematic. Therefore, in more recent report generators, more flexible structuring strategies are used. For instance, In BT-45, “key events” are always mentioned first. They are followed by events that are explicitly linked to the key event, which are in their turn followed by other co-temporal events. The messages on a key event with the messages on the events that are related to it form a paragraph in the report. The key event paragraphs are ordered by the start time of their respective key event. In LFS, statements are ordered according to their relative salience; a global salience

assessment procedure ensures that statements on the most significant economic changes are moved to the beginning of the text plan.

In MARQUIS, a mixture of schema-based planning at the paragraph level and top-down (Rhetorical Structure Theory) discourse structure-based planning within the individual paragraphs is realized [Bouayad-Agha et al., 2006] – which makes, as in TechDoc [Rösner and Stede, 1992], a prior (context- and user-tailored) identification of discourse relations that hold between statements necessary.

11.4 LINGUISTIC REALIZATION IN REPORT GENERATORS

The importance of the linguistic realization for RG is controversial. Thus, some report generators, such as TEMSIS, use templates, with gaps left open for words to be filled in. In SumTime-Mousam, SumTime-Turbine and BT-45, linguistic realization plays a minor role; some attention is given only to lexicalization and aggregation. On the other side, we observe rather elaborate linguistic modules in Gossip [Iordanskaja et al., 1991], FoG [Goldberg et al., 1994], LFS [Iordanskaja et al., 1992], Project Reporter [Korelsky et al., 1993], PLANDOC, and MARQUIS. McDonald [1993] argues that realization is a relatively trivial task in generation. In reality, the importance and complexity of the linguistic realization decisively depends on a variety of criteria into which we cannot delve here. It suffices to cite [Reiter, 1995] that a thorough cost-benefit analysis is needed to decide upon the most adequate linguistic realization. In PESCaDO, a powerful and flexible linguistic realization cannot be avoided.

11.4.1 INPUT AND LEVELS OF LINGUISTIC REALIZATION

The nature and degree of concretization of the text plan that serves as input to the linguistic realization module depends on the linguistic framework that underlies a report generator. It can be more or less abstract. Consider below the input structure to Streak's [Robin, 1994; Elhadad and Robin, 1996] linguistic module that is based on Systemic Functional Linguistics (SFL) [Halliday and Matthiessen, 1999, 2004]:

```
(cat clause
    (process (type material) (effect-type creative) (lex ``score'') (tense past)
    (participants ((agent (cat proper)
                          (head (cat person-name)
                                (first-name (lex ``Michael''))
                                (last-name (lex ``Jordan'')))))
                  (created (cat np) (cardinal (value 36))
                           (head (lex ``point'')) (definite no)))))
```

PLANDOC, which is also based on SFL, deals with similar input structures.

The linguistic models of report generators that are based on the Meaning- Text Theory, MTT, [Mel'cuk, 1998], among them Gossip, LFS, FoG, MultiMeteo, and MARQUIS, take as input either conceptual graphs in the sense of Sowa [Sowa, 2000] or semantic predicate-argument structures of the kind:

```
mean {sem= 'mean'
      1→ {index {sem=`index'
          1→ quality {sem=`quality'
                      1→air {sem='air'}}
          2→6 {sem=`six'}}
      2→ quality {sem=`quality'
          1→air {sem=`air'}
          2→ poor {sem=`poor'}}}
```

very {sem=`very'

1→poor}

Depending on the linguistic framework and the complexity of the linguistic phenomena of the domain, a message received as input to the linguistic realization module can undergo concretization in several stages. The number of stages (or levels of linguistic representation) varies. Thus, a number of generators (among them, e.g., TEMSIS and the SumTime generators) realize a direct projection of content structures to lexicalized syntactic structures. This does not exclude that the latter can be further modified, aggregated or paraphrased – but the level of abstraction remains always the same.

MTT distinguishes, apart from the semantic structure (SemS), a deep-syntactic (DSyntS), a surface-syntactic (SSyntS), a deep-morphological (DMorphS), and a surface-morphological structure (SMorphS). Often, a conceptual structure as an interface representation between content-oriented text plan codification and semantic message codification is added. Conceptual structures are language independent, while all the other structures are language-specific. But not all MTT-based generators use all of them. Thus, while MARQUIS uses indeed all, LFS unifies DMorphS and SMorphS, and FoG also skips SemS. In order to facilitate the choice between semantically equivalent but syntactically and lexically deviating variants (cf., e.g., *The high ozone concentration made the AQ index rise* vs. *The AQ index rose due to the high ozone concentration* in MARQUIS), some generators use the information (or, in the MTT terminology, communicative) structure. The information structure defines the distribution of information in a propositional structure in terms of oppositions such as Theme vs. Rheme, Given vs. New, etc. [Lambrecht, 1994; Mel'cuk, 2001]. Cf., [Iordanskaja et al., 1991 and Wanner et al., 2003] for illustration of the use of the information structure in RG. Current report generators that are not based on any formal linguistic framework tend to neglect the information structure, but keep track of the anaphoric structure which contains the co-reference links between units in order to be able to generate appropriate referring expressions by pronominalizing, choosing the definite article, or a hyperonym.

11.4.2 TASKS OF LINGUISTIC REALIZATION

In general discourse NLG, linguistic realization is often divided into the tasks of micro-(sentence) planning and surface realization. Microplanning deals with such tasks as sentence packaging, syntactic structure determination, aggregation, lexicalization, and referring expression generation. Surface realization is then a rather straightforward instantiation of the resulting structure of microplanning. In RG, only a subset of the microplanning tasks has been addressed so far. Most of sentence packaging is, as a rule, avoided in that each message is a priori considered a clause. Some generators use rule-based algorithms to merge either already available messages or clauses to obtain a more uid text (as, e.g., BT-45). In Streak and PLANDOC, special attention is given to lexicalization.

11.4.2.1 SYNTACTIC STRUCTURE DETERMINATION

A number of generators use syntactic structure templates on which the content structures are mapped directly – as, e.g., SumTime-Turbine. These templates can be rather abstract and formal (in the case of BT-45, in a HPSGlike format). Most MTT-based generators (e.g., Gossip, LFS, and MARQUIS) use the Theme/ Rheme distribution defined over the semantic (or conceptual) structures and the dominant or entry node of the latter to determine "deep-syntactic" structures. The dominant node, comparable to the key-event in BT-45, becomes the root of the syntactic tree; a Theme/Rheme partition embedded into another Theme or Rheme partition becomes a subordinate or relative clause.

11.4.2.2 AGGREGATION

Aggregation, i.e., fusion of partly overlapping structures with the purpose to avoid redundancies and achieve a more fluid text, can be carried out at different levels of a linguistic representation (as, e.g., in PLANDOC). Most often, however, it is done either prior to the linguistic realization proper at the message (content) level (as in BT-45) or during the transition between the semantic and syntactic level (as in MARQUIS). However, it is important to be aware

that aggregation can be semantic, syntactic or lexical [Dalianis, 1999], and that for more complex RG, a differentiation might be needed.

11.4.2.3 LEXICALIZATION

Lexicalization, i.e., mapping of content or semantic units onto lexical units (LUs), is traditionally given a more prominent role in RG. This is, on the one hand, due to the prominent place of lexicalization in general discourse NLG, and, on the other hand, due to the idiosyncratic vocabulary of the sublanguages in report domains. Four types of lexicalization need to be dealt with:

(i) full LUs, i.e., content words that correspond to one or several units in a message that is to be lexicalized;
(ii) collocation units such as heavy storm, dramatic loss, sharp rise, etc., where the choice of one of the LUs is contingent on the other LU;
(iii) functional words whose introduction is controlled by either subcategorization or idiosyncratic lexical restrictions of an LU, as, e.g., *relevant to* (and not **relevant for*) and *at* [*a*] *a measuring station* (and not **in* [*a*] *measuring station*);
(iv) discourse markers that connect messages and make explicit temporal, causal and other types of relations that hold between these messages.

Most of the report generators focus on the choice of full LUs. For instance, BT-45 maps messages onto case frame like representations with a verbal predication as the head and thematic roles as arguments. The lexicalization is straight-forwardly linked to the ontology. In SumTime-Mousam, the input to microplanning are tuples of the kind (0600, 8, 13, W, nil): time, wind speed lower range, wind speed higher range, wind direction, "modifier" (such as 'gust', 'shower'); each element in the tuple has a number of lexicalized phrase templates associated with it. In MARQUIS, the semantic dictionary gives for each semanteme all its possible lexicalisations. For instance, the meaning `cause' is mapped to the LUs cause[V], cause[N], result[V], result[N], due, because, consequence, etc.

Most of the MTT generators use the advanced lexicalization instrument offered by the theory – the lexical functions, LFs [Mel'cuk, 1996]. LFs are a formal means to encode idiosyncratic names and collocations in a generic way. In a functional notation, they are written as follows: Magn(*storm*) = *heavy*, Magn(*loss*) = *dramatic*, Oper1(*concentration*) = *have*, IncepFunc2(*concentration*) = *reach*, etc. The LFs of an LU may be referenced in the semantic dictionary for lexicalization (as in MARQUIS) or be chosen in a paraphrasing stage that follows the default lexicalization (as in LFS). With a few exceptions, as, e.g., ANA [Kukich, 1983], other report generators do not deal with collocations. ANA encodes them in a phrasal lexicon [Becker, 1975]. The problem of encoding phraseological information in a lexicon of this type is that it is hard to extend and maintain. That is, it might mean a rich variety of lexicalization in small domains (such as in stock market reports in the case of ANA), but can hardly sustain a large scale generator in an operational mode.

As the only report generator, Streak implements "reitified lexicalization": revision of lexical choices based, for instance, on aggregation criteria. Thus, having planned on the first pass two separate sentence structures, which would result, e.g., in *Karl Malone scored 39 points and Karl Malone's 39 point performance is equal to this season high*, respectively, Streak would conflate them in order to obtain *Karl Malone tied his season high with 39 points.*

11.5 MULTIMODAL RG

In the Introduction, we mentioned that the use of other modi than text has been largely neglected so far in RG. The only notable exception we know of has been MARQUIS. MARQUIS uses tables and graphics. However, both the corresponding generators and the strategies for the assignment of a specific mode to a specific content are very preliminary. The figure to the right displays a sample graphic generated by MARQUIS.

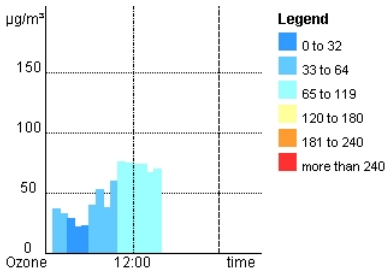


Figure 11.1: Use of the graphic mode in automatic report generation

The content-mode assignment is predefined and done at the discourse schema element level. For instance, the archive information is always provided in a table; the concentration of a primary pollutant substance during the last measurement is provided as text, etc.

11.6 The assessment of the state of the art for PESCaDO

The above review how state of the art report generators address the individual tasks reveals that in order to achieve the objectives of PESCaDO related to the delivery of multilingual and multimodal user-tailored information, the following themes must be addressed:

(a) Discourse planning techniques that do not rely on predefined discourse schemata, but are still able to take into account domain discourse restrictions and text genre restrictions that are to be generated: reports, instructions (suggestions, recommendations and indications), and warnings.

Current report generator planners are all domain- and even application-specific. A main scientific goal in PESCaDO must be to clearly separate domain-specific aspects of the task from "universal" aspects in order to ensure that the developed techniques are flexible enough to ensure their use in other environmental information applications and beyond

(b) Powerful mode selection techniques that are intelligent enough to be driven by the layout characteristics of the content and the preferences of the user (rather than by predefined ad hoc criteria).

(c) Multilingual techniques for purpose-driven linguistic generation of environmental information material, and, in particular, techniques for sentence planning techniques.

11.7 References

[Barzilay and Lapata, 2005] Barzilay, R. and M. Lapata. "Collective content selection for concept-to-text generation." In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 338, 2005.

[Becker, 1975] J.D. Becker. The Phrasal Lexicon. In R.C. Schank and B.L. Nash- Webber, editors, Theoretical Issues in Natural Language Processing (TINLAP): 1, pages 70-73. Bolt, Cambridge, MA, 1975.

[Binsted and Ritchie, 1996] K. Binsted and G. Ritchie. Speculations on Story Puns. In Proceedings of the International Workshop on Computational Humour, pages 151-159, Enschede, NL, 1996.

[Bohnet et al., 2007] B. Bohnet, F. Lareau, and L. Wanner. Automatic Production of Multilingual Environmental Information. In Proceedings of the EnviroInfo Conference, Warsaw, 2007.

[Bohnet et al., 2001] B. Bohnet and L. Wanner et al. Autotext-UIS: Automatische Produktion von Ozonkurzberichten im Umweltinformationssystem Baden-Wüurttemberg. In Proceedings of the Workshop Hypermedia und Umweltschutz, Ulm, 2001.

[Bontcheva, 1997] K. Bontcheva. Generation of Multilingual Explanations from Conceptual Graphs. In Recent Advances in Natural Language Processing, pages 365-376. Benjamins, Amsterdam, Philadelphia, 1997.

[Bontcheva and Wilks, 2004] K. Bontcheva and Y. Wilks. Automatic Generation from Ontologies: The MIAKT Approach. In Proceedings of the International Conference on Applications of Natural Language to Information Systems, pages 324- 335, Salford, 2004.

[Bouayad-Agha et al., 2006] N. Bouayad-Agha, L. Wanner, and D. Nickla_. Discourse Structuring of Dynamic Content. In Proceedings of the Spanish Conference on Computational Linguistics (SEPLN), Zaragoza, 2006.

[Busemann and Horacek, 1997] S. Busemann and H. Horacek. Generating Air-Quality Reports from Environmental Data. In Proceedings of the DFKI Workshop on Natural Language Generation, pages 15-21, Saarbrüucken, Germany, 1997.

[Cabré, 1994] M.T. Cabré. Terminology. Theory, Methods and Applications. Benjamins, Amsterdam, Philadelphia, 1998.

]Caldwell and Korelsky, 1994] D. Caldwell and T. Korelsky. Bilingual Generation of Job Descriptions from Quasiconceptual Forms. In Fourt Conference on Applied Natural Language Processing, pages 1-6, Stuttgart, Germany, 1994.

[Carcagno and Iordanskaja, 1992] D. Carcagno and I. Iordanskaja. Content Determination and Text Structuring: Two Interrelated Processes. In H. Horacek and M. Zock, editors, New Concepts in Natural Language Generation, pages 10-26. Pinter Publishers, London, 1992.

[Cheong and Young, 2006] Y.-G. Cheong and R.M. Young. A Computational Model of Narrative Generation for Suspense. In Proceedings of the AAAI 2006 Computational Aestetic Workshop, Boston, 2006.

[Coch, 1996] C. Coch. Overview of ALETHGEN. In Proceedings of the 8th International Workshop on Natural Language Generation, Volume 2, pages 25-28, Herstmonceux, 1996.

[Coch, 1998] J. Coch. Interactive Generation and Knowledge Administration in MultiMeteo.In Ninth International Workshop on Natural Language Generation, pages 300-303, Niagara-on-the-Lake, Ontario, Canada, 1998.

[Dalianis, 1999] H. Dalianis. Aggregation in Natural Language Generation. Computational Intelligence, 15(4):384-414, 1999.

[Duboue and McKeown, 2003] Duboue, P. A., and K. R. McKeown. Statistical Acquisition of Content Selection Rules. Computer Science Technical Report Series, 2003.

[Elhadad and Robin, 1996] M. Elhadad and J. Robin. An Overview of SURGE: A Reusable Comprehensive Syntactic Realization Component. Technical Report, Ben Gurion University in the Negev, 1996.

[Gervás, 2001] P. Gervás. Modeling Literary Style for Semi-Automatic Generation of Poetry. In Proceedings of the 8th International Conference on User Modeling, Sonthofen, Germany, 2001.

[Giarratano and Riley, 2005] J. Giarratano and G. Riley. Expert Systems: Principles and Programming. PWS Publishing Company, Boston, MA, 2005.

[Goldberg et al., 1994] E. Goldberg, N. Driedger, and R. Kittredge. Using Natural Language Processing to Produce Weather Forecasts. IEEE Expert, April 1994.

[Gotti, 2003] M. Gotti. Specialized Discourse. Linguistic Features and Changing Conventions. Peter Lang, Bern, Switzeland, 2003.

[Halliday and Matthiessen, 1999] M.A.K. Halliday and C.M.I.M Matthiessen. Construing Experience through Meaning: A Language-Based Approach to Cognition. Continuum, London, New York, 1999.

[Halliday and Matthiessen, 2004] M.A.K. Halliday and C.M.I.M Matthiessen. Introduction to Functional Grammar. Oxford University Press, Oxford, 2004.

[Harris, 2008] M.D. Harris. Building a Large-Scale Commercial NLG System for an EMR. In Proceedings of the International Natural Language Generation Conference, pages 157-160, Salt Fork, OH, 2008.

[Hovy, 1988] E.H. Hovy. Generating Natural Language under Pragmatic Constraints. Lawrence Erlbaum, Hillsdale, New Jersey, 1988.

[Iordanskaja and Polguère, 1988] L. Iordanskaja and A. Polguère. Semantic Processing for Text Generation. In Proceedings of the International Computer Science Conference, Hong Kong, 1988.

[Iordanskaja et al., 1992] L.N. Iordanskaja, M. Kim, R. Kittredge, B. Lavoie, and A. Polguère. Generation of Extended Bilingual Statistical Reports. In COLING-92, pages 1019-1022, Nantes, 1992.

[Iordanskaja et al., 1991] L.N. Iordanskaja, R. Kittredge, and A. Polguère. Lexical Selection and Paraphrase in a Meaning-Text Generation Model. In C.L. Paris, W.R. Swartout, and W.C. Mann, editors, Natural Language Generation in Artifcial Intelligence and Computational Linguistics. Kluwer Academic Publishers, 1991.

[Kelly et al., 2009] Kelly, C., A. Copestake, and N. Karamanis. “Investigating content selection for language generation using machine learning.” In Proceedings of the 12th European Workshop on Natural Language Generation, 130–137, 2009.

[Kim et al., 2002] S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt, and M. Weal. Antequakt: Generating Tailored Biographies with Automatically Annotated Fragments from the Web. In Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop at ECAI 2002, Lyon, 2002.

[Kittredge and Lavoie, 1998] R. Kittredge and B. Lavoie. MeteoCogent: A Knowledge-Based Tool for Generating Weather Forecast Texts. In Proceedings of the American Meteorological Society AI Conference (AMS-98), Phoenix, Arizona, 1998.

[Kittredge and Lehrberger, 1982] R. Kittredge and J. Lehrberger, editors. Sublanguage: Studies of Language in Restricted Semantic Domains. de Gruyter, Berlin, 1982.

[Kittredge et al., 1986] R. Kittredge, A. Polguère, and E. Goldberg. SynthesizingWeather Forecasts from Formatted Data. In Proceedings of the Computational Linguistics Conference (COLING) '86, pages 563-565, Bonn, 1986.

[Kittredge and Polguère, 2000] R.I. Kittredge and A. Polguère. The Generation of Reports from Databases. In R. Dale, H. Moisl, and H. Somers, editors, Handbook of Natural Language Processing, pages 261{304. Taylor and Francis, New York, 2000.

[Korelsky et al., 1993] T. Korelsky, D. McCullough, and O. Rambow. Knowledge Requirements for the Automatic Generation of Project Management Reports. In Proceedings of the Eighth Knowledge-Based Software Engineering Conference, pages 2-9. IEEE Computer Society Press, 1993.

[Kukich 1993] K. Kukich. Knowledge-Based Report Generation: A Technique for Automatically Generating Natural Language Reports from Databases. In Proceedings of the Sixth International ACM SIGIR Conference, Washington DC, 1983.

[Lambrecht, 1994] K. Lambrecht. Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents. Cambridge Studies in Linguistics 71. Cambridge University Press, Cambridge, 1994. References 31

[Lu et al., 2000] S. Lu, F. Paradis, C. Paris, S. Wan, R. Wilkinson, and M. Wu. Generating Personal Travel Guides from Discourse Plans. In Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems, 2000.

[Maybury, 1990] M. Maybury. Using Discourse Focus, Temporal Focus, and Spatial Focus to Generate Multisentential Text. In Proceedings of the 5th International Workshop on Natural Language Generation, pages 70{78, Dawson, PA, 1990.

[McDonald, 1993] D. McDonald. Issues in the Choice of a Source for Natural Language Generation. Computational Linguistics, 19:191-197, 1993.

[McKeown, 1985] K. McKeown. Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Cambridge University Press, Cambridge, England, 1985.

[McKeown et al., 1994] K. McKeown, K. Kukich, and J. Shaw. Practical Issues in Automatic Documentation Generation. In Proceedings of the Fourth Conference on Applied Natural Language Processing, pages 7{14, Stuttgart, Germany, 1994.

[McKeown et al., 1995] K. McKeown, J. Robin, and K. Kukich. Generating Concise Natural Language Summaries. Information Processing and Management, 31:703-733, 1995.

[Mel’cuk, 1988] I.A. Mel'cuk. Dependency Syntax: Theory and Practice. SUNY Press, Albany, 1988.

[Mel’cuk, 1996] I.A. Mel'cuk. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner, editor, Lexical Functions in Lexicography and Natural Language Processing, pages 37-102. Benjamins Academic Publishers, Amsterdam/Philadelphia, 1996.

[Mel’cuk, 2001] Igor A. Mel'cuk. Communicative Organization in Natural Language (The Semantic-Communicative Structure of Sentences). Benjamins Academic Publishers, Amsterdam, 2001.

[Mellish and Dale, 1998] Mellish, C., and R. Dale. “Evaluation in the conext of natural language generation.” Computer Speech and Language 12, no. 4 (1998): 349–374.

[Paris, 1993] C. Paris. User Modelling in Text Generation. Frances Pinter Publishers, London, 1993.

[Paris et al., 1995] C. Paris, K. Vander Linden, M. Fischer, A. Hartley, L. Pemberton, R. Power, and D. Scott. A Support Tool for Writing Multilingual Instructions. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, 1995.

[Portet, 2009] F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. Artificial Intelligence, 173(7-8):789-916, 2009.

[Rambow, 1990] Owen Rambow. Domain Communication Knowledge. In Proceedings of the 5th Natural Language Generation Workshop, June 1990, pages 87-94, Dawson, PA., 1990.

[Reiter, 1995] E. Reiter. NLG vs. Templates. In Proceedings of the 5th European Workshop on Natural Language Generation, pages 95-104, 1995.

[Reiter, 2007] E. Reiter. An Architecture for Data-to-Text Systems. In Proceedings of the 11th European Workshop on Natural Language Generation, pages 97-104, 2007.

[Reiter and Dale, 2000] E. Reiter and R. Dale. Building Natural Language Generation Systems. Cambridge University Press, Cambridge, 2000.

[Reiter et al., 2003] E. Reiter, R. Robertson, and L. Osman. Lessons from a Failure: Generating Tailored Smoking Cessation Letters. Artificial Intelligence, 144:41-58, 2003.

[Robin, 1994] J. Robin. Revision-Based Generation of Natural Language Summaries Providing Historical Background. PhD Thesis, Graduate School of Arts and Sciences, Columbia University, New York, 1994.

[Rösner, 1986] D. Rösner. Ein System zur Generierung von deutschen Texten aus semantischen Repräsentationen. Ph.D. Thesis, Institut für Informatik, Stuttgart University, Stuttgart, Germany, 1986.

[Rösner and Stede, 1992] D. Rösner and M. Stede. Customizing RST for the Automatic Production of Technical Manuals. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, Aspects of Automated Natural Language Generation. Springer Verlag, Berlin, 1992.

[Rösner and Stede, 1994] D. Rösner and M. Stede. Generating Multilingual Documents from a Knowledge Base: The TechDoc project. In Proceedings of COLING-94, pages 339-346, 1994.

[Sowa, 2000] J. Sowa. Knowledge Representation. Brooks Cole, Pacific Grove, CA, 2000.

[Sripada et al., 2003] S. Sripada, E. Reiter, and I. Davy. SumTime-Mousam: Configurable Marine Weather Forecast Generator. Expert Update, 6(3):4-10, 2003.

[Szilas, 2000] N. Szilas. A Computational Model of an Intelligent Narrator for Interactive Narratives. Applied Artificial Intelligence, 21(8):753-801, 2000.

[Wanner, 2010] L. Wanner. Report Generation. In N. Indurkhya and F. Damerau (eds.) Handbook of Natural Language Processing. CRC Press, Taylor and Francis, London, 533-556, 2010.

[Wanner, 1996] L. Wanner. Lexical Choice in Text Generation and Machine Translation. Machine Translation, 11(1-3):3-35, 1996.

[Wanner et al., 2007] L.Wanner, D. Nicklass, B. Bohnet, N. Bouayad-Agha, J. Bronder, F. Ferreira, R. Friedrich, A. Karppinnen, F. Lareau, A. Lohmeyer, A. Panighi, S. Parisio, H. Scheu-Hachtel, and J. Serpa. From Measurement Data to Environmental Information: MARQUIS-A Multimodal Air Quality Information Service for the General Public. In A. Swayne and J. Hrebicek (eds.), Proceedings of the 6th International Symposium on Environmental Software Systems, Prague, 2007.

[White and Caldwell, 1998] M. White and T. Caldwell. EXEMPLARS: A Practical, Extensible Framework for Dynamic Text Generation. In Proceedings of the Ninth International Natural Language Generation Workshop, pages 266-275, Niagara-on-the-Lake, Ontario, 1998.

[Williams et al., 2003] S. Williams, E. Reiter, and L. Osman. Experiments with Discourse-Level Choices and Readability. In Proceedings of the 9th European Natural Language Generation Workshop at the 10th Conference of the EACL, pages 127-134, 2003.

[Yao et al., 1998] T. Yao, D. Zhang, and Q. Wang. MLWFA: Multilingual Weather Forecasting System. In Proceedings of the Ninth International Natural Language Generation Workshop, pages 296-299, Niagara-on-the-Lake, Ontario, 1998.

[Yu et al., 2007] J. Yu, E. Reiter, J. Hunter, and C. Mellish. Choosing the Content of Textual Summaries of Large Time-Series Data Sets. Natural Language Engineering, 13(1):25-49, 2007.

[Zukerman and Litman, 2001] I. Zukerman and D. Litman. Natural Language Processing and User Modeling: Synergies and Limitations. User Modeling and User-Adapted Interaction, 11:129-158, 2001.

12 CHALLENGES FOR PESCADO IN THE LIGHT OF THE STATE OF THE ART

The analysis of the state of the art in both the design and orientation of environmental services as a whole and the related individual research areas revealed that the work in PESCaDO can draw upon previous work, but that this work is by far not sufficient to accomplish the objectives of the project. Thus, as far as environmental services as a whole are concerned, they are still very much conceived as one-way information delivery, with the system as a broadcaster and the end user as a passive consumer. No or only little interaction between the system and the user takes place, and only in a few prototypical services (such as MARQUIS and APNEE) personalization of the information delivery is foreseen. A high degree of personalization is achieved in APNEE for users using the mobile phone as information access device in that the location of the user is tracked such that location-relevant information can be supplied. None of the services considered so far the option to complement its own information sources with further potentially useful sources or to calibrate its data with competing data sources. In meteorological, road weather, water quality and similar services, the information is commonly presented in terms of pictograms (enriched by small text snippets), in air quality services, graphical curves still dominate. However, while pictograms are appropriate for overviews (also of air quality; cf. APNEE), personalized and more detailed information requires the use of report generation technologies (as exploited in MARQUIS).

To summarize, in order to realize its view of a personalized, active and collaborative environmental service, PESCaDO will thus have to change the predominating image of environmental services. But what is the situation with respect to the individual areas involved in PESCaDO? Let us briefly assess this in what follows.

As pointed out in Section 3, the discovery of environmental service nodes in the Web can be considered a problem of domain-specific Web search and web service discovery. Although the research on domain-specific search engines is well-established, there are no environment-domain search engines available so far. Web service discovery is nowadays mostly performed with the aid of register portals. Also, the use of existing search engines combined either with keyword spices or analysis result techniques incorporates the disadvantages of web search engines such as the inability to identify nodes in the Invisible Web and although web services provide well-defined and organized information, if their discovery is based mainly on portals, where the developers register voluntary their services, only part of the services can be acquired. Therefore, we can deduce that neither of the aforementioned techniques can fully cover the discovery of environmental nodes in the Web, when used separately. However, it is to be expected that an effective combination of the existing techniques, with the appropriate optimization and tuning of the employed modules, can aggregate their advantages and thus provide an adequate coverage of the information dispersed in the Web. For instance, a combination of the keyword-spice technique with crawler techniques in order to obtain non-registered environmental web services is expected to improve the results.

The uncertainty metrics (discussed in Section 4) pose a serious challenge to state-of-the art environmental services. PESCaDO will thus have to actively research reliable high quality metrics that are suitable for orchestration and the choice of the best data sources for the end user. In this context, PESCaDO will also seek collaboration with related initiatives.

All protocols and services described in Section 5 are relevant to PESCaDO. In the course of the design and implementation of the PESCaDO architecture, the contribution of PESCaDO in this area will be detailed.

The discussion of the state of the art in the area of the orchestration of environmental services (Section 6) made clear that so far the initiative and the process of orchestration is left to the end user: it is the end user who has to search for the adequate services that match (i.e., that are to be orchestrated), to figure out how to connect them, etc. PESCaDO will have thus do ground-breaking work in this area.

The assessment of the availability of environmental ontologies (Section 7) makes clear that, as expected, quite a few ontologies already exist and can thus be used in PESCaDO. Obviously, at this stage it is impossible to estimate how complete these ontologies are with respect to the knowledge to be covered in PESCaDO. What seems obvious is

that PESCaDO will have to align the existing ontologies in order to take advantage of as many of them as possible and to be prepared to extend the obtained ontology configurations. For this purpose, we can also draw upon technologies that have been developed in the context of ontology research. However, this still leaves the question of the extraction (or distillation) of knowledge from PESCaDO relevant sources (cf. Section 8). So far, no content distillation techniques have been tested with multilingual environmental material. As Section 8 points out, this is likely to be less critical for English since for English stable tools and external resources that support content distillation are available. In contrast, for Swedish and, even more so, for Finnish, the tools and resources are scarce. PESCaDO will thus need to research the necessary means for these two languages.

A further theme that PESCaDO will need to tackle in the context of content distillation is the extraction of content from multimodal material, notably from graphics and pictograms that play an important role in the material to be digested by PESCaDO.

As far as the state of the art in reasoning and inference techniques for user-oriented decision support is concerned (Section 9), we can conclude that already quite a few stable high quality reasoners that operate on semantic resources are available and can thus be used in PESCaDO. Reasoning on uncertain data has also been tackled in the past, but to a much lesser extent. PESCaDO sees thus itself contributing significantly in this area to be able to achieve its objectives.

The user-system interaction techniques in the context of environmental services did not receive due attention as yet (Section 10). Thus, to the best of our knowledge, no visual query formulation approach for personalized decision support in the domain of environmental services exists. In the scope of the PATExpert project, a hybrid and visual query formulation approach has been adopted. It is reasonable for PESCaDO to start with the adaptation of the PATExpert approach to decision support in the environmental service domain – although prior to any adaptation activities, the suitability of this approach must be examined. PESCaDO will also have to develop visual metaphors for the Problem Description Language (PDL) building blocks which need to be precise and flexible at the same time. To close the loop between request and result, novel techniques for interacting with the result representation will be developed.

The Visual Analytics support for the uncertainty metric determination process needs research of available data in the corresponding domain. Here, we need a close collaboration with the development of the metric model to be able to realize an on-line adjustable metric model that can deliver feedback at rates that are suitable for interactive visualization of the current metric configuration. Developing a representation of the metric that allows direct manipulation of its parameters as well as an expressive and visually comprehensible representation of the outcome of the metric will be the main challenge for the visualization module in PESCaDO.

The review of the techniques for the generation and delivery of information (Section 11) shows that PESCaDO can draw upon existing works – especially on the report generator developed in the MARQUIS project. However, the review also revealed significant shortcomings in the areas of multimodal discourse planning, sentence planning. These two topics will be the two major challenges in PESCaDO. In addition, generation resources for the three languages involved in PESCaDO (English, Finnish and Swedish) must be created.