PESCaDO

FP7-248594

Personalized Environmental Service Configuration and Delivery Orchestration

D4.1 Inventory of Environmental Ontologies and Corpora from the Environmental Domain

Due date of deliverable: 30.06.2010
Actual submission date: 30.06.2010

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

Lead contractor for this deliverable: FBK <Revision <Version>>

D4.1	Inventory of Environmental Ontologies and Corpora from the Environmental Domain
Project Acronym :	PESCaDO
Contract No :	FP7 - 248594
Due Date :	30.06.2010
Reply To:	Marco Rospocher rospocher@fbk.eu
Actual date of delivery:	30.06.2010

Deliverable Identification Sheet

Project ref. no.	FP7-248594
Project acronym	PESCaDO
Project full title	Personalized Environmental Service Configuration and Delivery Orchestration
Security (distribution level)	PU
Contractual date of delivery	Month 6, 30.06.2010
Actual date of delivery	Month 6, 30.06.2010
Deliverable number	D4.1
Deliverable name	Inventory of Environmental Ontologies and Corpora from the Environmental Domain
Type	Report
Status & version	Submitted, V1
Number of pages	24
WP / Task responsible	FBK
Other contributors	
Author(s)	Emanuele Pianta, Marco Rospocher, Luciano Serafini, Sara Tonelli
Internal Reviewer	Stefanos Vrochidis, Leo Wanner
EC Project Officer	Manuel Monteiro
Abstract	This deliverable gives a detailed insight of existing ontologies and corpora covering several environment-related topics such as meteorological conditions, air quality, pollen information, traffic situation, human health and geospatial information. Besides, it describes some general strategies to acquire/integrate such resources in case of gaps.

Table of Contents

Executive Summary

The present document provides a description of the most relevant ontologies and corpora for the environmental domain. It is divided into two main sections. The first part is devoted to the presentation of the study of the state of the art on environmental ontologies. First, we recall the main motivations for the use of ontologies in the project. Then, we introduce and briefly describe the domains to be covered by ontologies in PESCaDO. For each of the domains considered, we report our analysis of existing ontologies covering it. Finally, we conclude with some considerations regarding domains for which suitable ontologies have not been found.

The second part focuses on the description of a collection of environmental documents created to build a domain and language-specific corpus. In particular, a list of the available documents issued by official institutions in the field is given, as well as a description of environmental websites from which the content in plain text is being automatically extracted to build the corpora for the project. Finally, we briefly discuss how environmental corpora can be used in PESCaDO in connection with the above mentioned ontologies. For both types of resources, the available material will need further enrichment and refinement: for example, the existing ontologies will have to be connected and the resulting gaps will have to be eliminated by adding the missing concepts. As for the corpora, all relevant domains will have to be covered by English, Swedish and Finnish texts, possibly through the creation of parallel or comparable document collections.

1 Environmental Ontologies

1.1 Purpose of ontologies in PESCaDO

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, will be founded on an ontological representation of the environmental domain, in order to guarantee:

- *semantic orchestration of heterogeneous environmental service nodes*: PESCaDO will investigate the applicability of ontology-based techniques for the selection of
  a) competitive nodes for mutual data validation and for derivation of more reliable common data from the outputs of the individual nodes,
  b) complementary nodes for completion of the data provided by individual nodes, and
  c) to-be-chained-in nodes that provide data required for other nodes as input;
- *user-oriented decision support*: to provide decision support to the user, PESCaDO will implement reasoning techniques and strategies based on the information stored in the ontologies;
- *environmental information delivery*: PESCaDO will develop user-tailored multilingual environmental information generation techniques that offer information to the users either to brief them or to verbalize advices and suggestions deduced by reasoning and interpretation mechanisms.

This ontological representation of the environmental domain has to be as much as possible comprehensive, in order to adequately cover the domain and, in particular, the PESCaDO use cases domains.

Although a comprehensive ontological representation of the environmental domain as needed in PESCaDO is not currently available among the state of the art resources, it is unlikely that it will have to be built entirely from scratch. In fact, some 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 similar entities may be described in more than one ontology. Furthermore, parts of the knowledge of the environmental domains relevant to PESCaDO may not be covered by any available ontology.

In this chapter, we first introduce and briefly describe some environmental related domains that we recognize as relevant for PESCaDO. Then, we report about available ontologies covering the domains considered, concluding with some final considerations.

1.2 Domains to be covered by PESCaDO ontologies

We made a first selection of the environmental related domains that have to be covered by the ontologies upon which the PESCaDO workbench will rely on. This first selection has been performed taking in consideration the preliminary specification of the Pilot Use Cases in PESCaDO (which will be documented in details in Deliverable D8.5). The first use case addresses ordinary (non-professional) citizens, helping them in planning a journey or doing an activity. It comprises environmental issues affecting the decision of journey e.g. weather, air quality, pollen, and traffic aspects. The system is supposed to give an answer to questions like "Are there any health issues if I travel tomorrow from X to Y?". The users addressed in the second use case covers experts and administrative clients of the services offered by HSY for the Helsinki Metropolitan Area (HMA). It concentrates on (administrative) decision support by combining information from different branches, e.g. environmental (air quaility and weather) observed and forecasted data, administrative action plans, the information of other administrators etc. Other aspects we considered in the selection of the domain are the environmental data typically available at environmental related sites, and the basic environmental related terminology/concepts.

More specifically, we target the following environmental related domains:

- Meteorological Conditions and Phenomena
- Air Quality
- Pollens

In addition to the above strictly environmental domains, knowledge about the following ones will be relevant for PESCaDO in order to provide adequate user decision and suggestion support:

- Travel and traffic information
- Human Health
- Geospatial Information and geographic datasets

These domains will be described in detail in the following sections.

1.2.1 Meteorological Conditions and Phenomena

The meteorological conditions and phenomena domain covers typical concepts and attributes commonly found in weather reports and weather forecasts. As such, relevant concepts that are characteristic of this domain are:

- temperature, temperature value, and units of measurement of temperatures;
- type of precipitations, including e.g. rain, snow, hail, etc.;
- intensity and types of wind, like e.g. MountainWind, SeaWind, Foehn, Maestro, etc.;
- sky conditions, like e.g. cloudy, partly cloudy, sunny, etc.;
- sun exposure and UV index;
- atmospheric pressure and units of measurement of atmospheric pressure.

1.2.2 Air Quality

This domain covers typical concepts and attributes dealing with the quality of air. As such, particularly relevant for this domain are concepts related to:

- air quality terminology, air quality indexes (AQI) and air quality levels (e.g. Good, Satisfactory, Fair, Poor, Very Poor);
- air pollutants monitored and commonly found in the air, like e.g. ParticulateMatter$_{10}$, Ozone, etc., and units of measurement of air pollutants concentrations.

According to the use cases considered, it should be noticed that the focus of this domain is on *Outdoor* Air Quality, hence knowledge specifically about *Indoor* Air Quality will not be considered relevant.

1.2.3 Pollen

This domain covers the typical concepts and attributes dealing with pollens and pollen counts (i.e. the concentration of a pollen in the air). As such, relevant for this domain are concepts related to:

- pollen types commonly monitored (e.g. grass pollens, birch pollen), and units of measurement of concentrations of pollens in the air.

1.2.4 Travel and traffic information

This domain is about the travel and traffic information relevant from an environmental point of view. In particular, the relevant knowledge here considered will be about:

- Travels means (e.g. car, bus, train);
- Weather-related road conditions, i.e. the conditions of a road depending on the weather phenomena (e.g. slippery road conditions due to black ice or rain, road covered by snow);
- Road traffic situation, i.e. the situation of a road depending on the traffic load (e.g. reasonably free flow, unstable flow, traffic jam);
- Road situation with respect to on-going roadworks (e.g. road closed for resurfaced).

1.2.5 Human Health

The knowledge of domain is about human diseases and symptoms. In particular, environment affected and caused diseases, and environment triggered symptoms are highly relevant to PESCaDO. As such, terminology that needs to be represented about this domain will cover:

- Environmental related diseases, including respiratory system diseases (e.g. asthma, bronchitis, sinusitis), circulatory system diseases (e.g. hypertension, hypotension), skin and subcutaneous tissue diseases (e.g. dermatitis), etc.;
- Environmental related symptoms, including circulatory and respiratory systems symptoms (e.g. tachycardia, cough, chest pain), skin and subcutaneous tissue symptoms (e.g. rash, desquamation), etc.

1.2.6 Geospatial Information and geographic datasets

We are also currently considering describing with the aid of ontologies the knowledge about the geographical location of places, and relative geographical relations between places. The relevant concepts for this domain are considered to be related to:

- geographical coordinates of a location;

- Subdivision of geographical areas in sub-regions, or neighbourhoods of a certain area.

Note that, since the knowledge about this domain is quite static (i.e. it does not change very often over time), an ontology describing this domain could be directly setup with factual knowledge from available geographic datasets.

1.3 Inventory of environmental and environment-related ontologies

The state of the art on environmental ontologies includes some candidates that may be relevant to the domains of interest of PESCaDO, and thus may be utilized either in their initial format or partly adapted for inclusion in the PESCaDO ontology repository. Below we describe the most prominent candidates we selected, presenting them organized according to the domains mentioned in the previous section. Our selection is restricted to ontologies expressed in OWL, since this is nowadays the standard Semantic Web language for representing ontologies, and this will be the ontology language that will be used in PESCaDO.

The search of available ontologies has been performed by:

- looking at previously known environmental ontologies or environmental institutions offering ontological resources, like in the case of the SWEET;
- submitting queries to search engines like Swoogle (http://swoogle.umbc.edu/) and OntoSelect (http://olp.dfki.de/ontoselect/), which are specialised in retrieving ontologies;
- submitting queries to standard web search engines like Google (www.google.com/).

All the ontologies reported have been validated with Protégé 4.0 OWL, and consistency checks have been run with two state-of-the-art reasoners, Pellet and FACT++.

1.3.1 Meteorological Conditions and Phenomena ontologies

Regarding the meteorological conditions and phenomena domain, the following ontologies have been identified as relevant to PESCaDO.

SWEET (Semantic Web for Earth and Environmental Terminology)

Developed by: Propulsion Laboratory of California Institute of Technology
Url: http://sweet.jpl.nasa.gov/2.0/

SWEET is an ontology developed by the Jet Propulsion Laboratory of California Institute of Technology, which describes relevant knowledge in Earth Science and Environmental domains. The ontology is structured in modules arranged hierarchically by subject, as depicted in the figure 1 below.

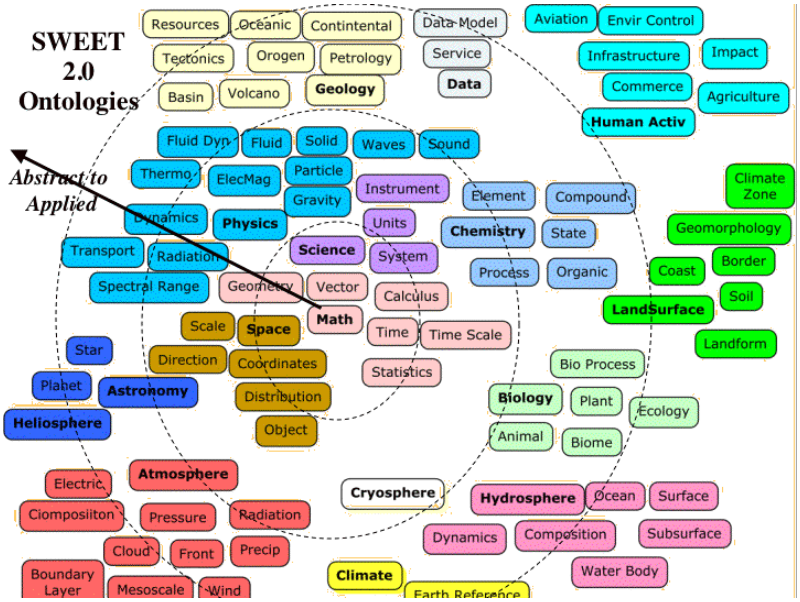

Figure 1 - Subjects covered by the SWEET ontology (taken from http://sweet.jpl.nasa.gov)

The latest available release at the time of writing this document is version 2.0 Beta, which consists of 150 ontology modules, and includes:

- 4953 concepts;
- 259 object properties;
- 42 data-type properties;
- 605 individuals.

The DL expressivity of the whole ontology is *SHOIN(D)*. The ontology passed successfully the consistency check.

Being quite a big ontology, it will be preferable, also from a computational point of view, to consider only those ontology modules that are relevant to the PESCaDO domains, instead of using the whole ontology. In particular, when considering only meteorological conditions and phenomena, the relevant modules of the SWEET ontology are those falling in the sub-hierarchy of

- the *Atmosphere* module, which describe a wide range of meteorological phenomena;
- the *PhysicProperty* module, which contains the description of thermodynamic properties, including e.g. temperature and windspeed;
- the *ScienceResearch* module, which contains the description of the main units of measurement.

In detail, the *Atmosphere* related sub-modules characterize many relevant meteorological phenomena and conditions. As an example, we show in Figure 2 the taxonomy describing sky conditions:

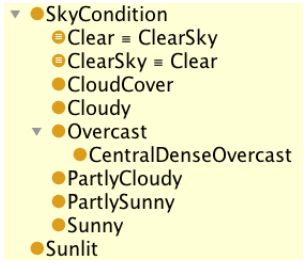

Figure 2 - Excerpt of the SkyConditions Taxonomy

Other meteorological phenomena and conditions described are atmospheric phenomena (which include a detailed classification of winds), anticyclones, atmospheric circulation, lightning, fronts, cloud types, etc. Particularly detailed is the classification of precipitations, an excerpt of which (limited to two hierarchical levels) is shown in the figure 3 below:

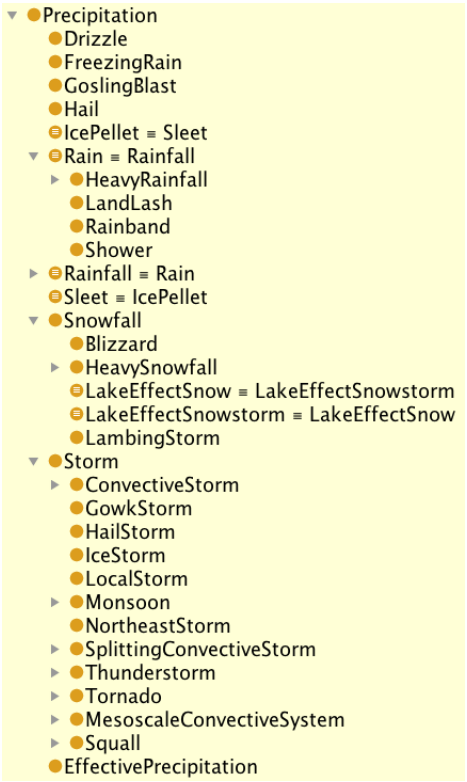

Figure 3 - Excerpt of the Precipitations taxonomy

Analogously, the *PhysicProperty* related sub-modules describe, among the others, many relevant environmental concepts, including atmospheric pressure, temperature, and wind speed.

Finally, the *ScienceResearch* related sub-modules contain, among the others, a description of many units of measurement relevant to the environmental domain, like e.g. Celsius degrees or Fahrenheit degrees, and environmental indicators, like the UVindex.

Weather Ontology

Developed by: WeatherBot project (Aaron Elkiss)
Url: http://www.csd.abdn.ac.uk/research/AgentCities/WeatherAgent/index.php[1]

Another available ontology is the Weather Ontology, developed within the WeatherBot project, and currently hosted and revised by the University of Aberdeen. The Weather Ontology has been built for converting METAR and TAF reports, which are considered as two standard formats for reporting weather forecast information, to DAML, which is a DARPA developed markup language for the Semantic Web. The ontology is also used by an agent system developed by the University of Aberdeen to report the current weather situation in Aberdeen, as well as a forecast a few days ahead.

The latest available release at the time of writing this document is version 1.12 (dated 23/01/2002), which includes:

- 98 concepts;
- 38 object properties;
- 1 data-type property;
- 38 individuals.

[1] The url of the WeatherBot project (http://dormouse.cs.umd.edu:8080/wiki/weatherbot.wiki) was not working at the time of writing this deliverable. However, a copy of the ontology is currently hosted by the University of Aberdeen at the link posted here.

The DL expressivity of the whole ontology is *ALUHN(D)*. The ontology passed successfully the consistency check.

Albeit being smaller in size with respect to the SWEET ontology, it is very focused on describing weather reporting related entities and properties, elements which are relevant for PESCaDO. As an example, it describes:

- concepts like e.g. *SkyCondition*, *CloudType*, *MaximumTemperature*, and *PrecipitationEvent*;
- object properties like e.g. temperatureMeasurement, windDirection and windSpeed.

Here below, we show an excerpt of the taxonomy of concepts (limited to the first two hierarchical levels) described in the Weather Ontology:

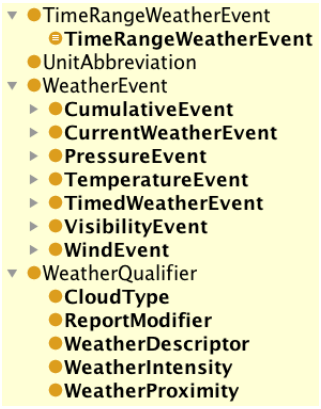

Figure 4 - Excerpt of the Weather Ontology taxonomy

The ontology is written in DAML+OIL, but an OWL version can be easily obtained thanks to some state of the art DAML-2-OWL converters[2].

MetBroker Ontology

Developed by: National Agricultural Research Center, Tsukuba, Japan
Url: http://www.agmodel.org/vocabulary/200608/MetBroker.owl

A third available ontology describing meteorological conditions & phenomena is MetBroker Ontology, developed by the National Agricultural Research Center in Japan. MetBroker is a webservice that focuses on providing agricultural Decision Support Systems with direct access to the kind of meteorological data that they require. To provide this access, the MetBroker ontology is used to virtually integrate distributed Meteorological Data.

At the time of writing this document, the latest available release of this ontology is an unnumbered version dated August 2006, which includes:

- 40 concepts;
- 9 object properties;
- 15 data-type properties;
- 20 individuals.

The DL expressivity of the whole ontology is *ALCO(D)*. The ontology passed successfully the consistency check.

Due to the purpose of the application for which at has been built, rather than providing a classification of precipitation or winds, like for example in SWEET Ontology, this ontology describe mostly quantitative properties of meteorological phenomena. Hence, for example,

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

concepts like SnowDepth or RainOneHourMax are defined. An excerpt of the taxonomy of concepts of MetBroker Ontology is described in the figure below:

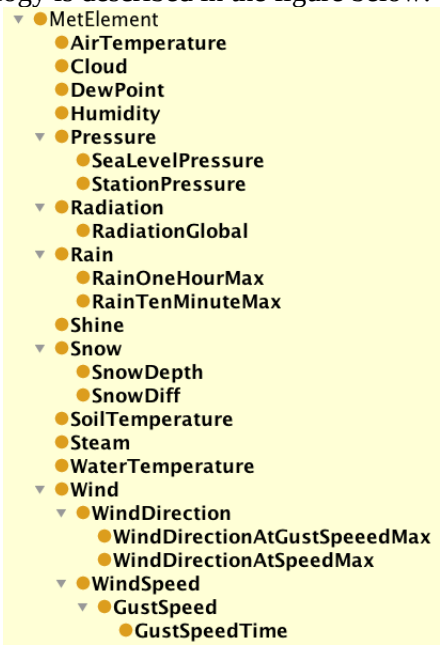

Figure 5 - Excerpt of the MetBroker Ontology taxonomy

1.3.2 Air Quality ontologies

Regarding the air quality and air pollutants domain, **SWEET Ontology** provides a description of many relevant concepts and properties, collected under the modules (and related sub-modules) *envirImpact*, *envirIndicator*, and *chemical*.

In module *envirImpact*, there is a terminological description of what is meant for AirPollution and AirQuality. In figure 6 we report an excerpt of the corresponding taxonomy:

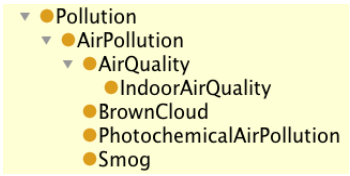

Figure 6 - Air Quality concepts in SWEET Ontology

Module *envirIndicator* defines the main relevant environmental related indexes. Among them, some relevant air quality indexes are defined. In figure 7 below we report an excerpt of the corresponding taxonomy:

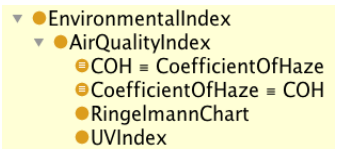

Figure 7 - Air Quality indexes in SWEET Ontology

It should be noticed that the air quality indexes described in SWEET ontologies are not specialised according to concentration levels. This is due to the fact that air quality levels

usually vary depending on the country considered. Thus, it will be necessary to extend this ontology with specific information depending on the country considered in our use cases, i.e. Finland.

Finally, the *chemical* modules describe many substances and air pollutants commonly monitored and regulated. Among them, particulate matter (both $PM_{2.5}$ and PM_{10}), ozone (O_3), nitrogen oxides (both NO_2 and NO), sulphur dioxide (SO2), etc are of importance

Clearly, an ontology for the air quality/pollution domain should at least contain a description of the air pollutants relevant to the Country considered in the use cases, that is Finland. At a first analysis, most of the air pollutants monitored in Finland are considered in SWEET ontology, but some are missing, like e.g. total reduced sulphur (TRS) (see e.g. http://www.ilmanlaatu.fi/ilmansaasteet/indeksi/indeksi.php).

An ontology developed specifically for the air quality/pollution domain, called AIR_POLLUTION_Onto, is also mentioned is some papers[3], however the corresponding OWL file is not available on the web.

1.3.3 Pollen ontologies

To the best of our knowledge, no specific ontology describing pollen types and pollens concentrations is currently available. Hence, this ontology will probably have to be built from scratch, starting from available resources, and also taking in consideration the PESCaDO use cases and the kind of knowledge required to be modelled.

Resources describing pollens and pollens concentration can be found at http://www.polleninfo.org, which, for most of the European region countries, reports:

- Pollen forecasts of upcoming pollen levels;
- Countdown to the beginning/ending of a specific pollen season;
- Flow charts reporting the mean pollen counts for a certain region;
- Distribution maps showing, for each specific pollen, the intensity load over Europe.

Particularly relevant from an ontological prospective is the POLLENATLAS (http://www.polleninfo.org/index.php?language=en&nav=_n2&module=article&action=first_page&row=0&id_parent=2055), which contains pictures and descriptions of nearly 100 pollen types that are found in air samples, categorized according to criteria like specie and family.

The main pollens currently monitored in Finland are Alder, Birch, Grasses, and Mugwort (as reported here http://aerobiologia.utu.fi/tiedotus/siitepolytiedote/polleninformation.html), and the information is based on air particles count monitored at nine Finnish locations. Other pollens monitored are ash tree, cypress family, hazel, and oak (see http://www.polleninfo.org/index.php?language=en&nav=&module=states&action=first_page&row=9&id_parent=&id_parent=&id_parent=9&register=_r4).

1.3.4 Travel and traffic information

To the best of our knowledge, no specific ontology describing comprehensively the travel and traffic information domain is currently available, at least for what concerns the aspects relevant for PESCaDO.

The development of a prototype ontology for the semantic modelling of road traffic information elements is presented in a conference paper[4], in the context of the CICYT project[5]. The purpose

[3] Mihaela M. Oprea – "AIR_POLLUTION_Onto: an Ontology for Air Pollution Analysis and Control" – Artificial Intelligence Applications and Innovations III, Volume 296/2009, pp135-143.

of the ontology is to describe relevant traffic-related domains, like e.g. classification of roads (e.g. motorway), vehicles (e.g. truck), traffic events (e.g. an accident), people roles (e.g. driver), and routes (e.g. urban). However, to best of our knowledge, only some modules on the typology of land transportations (according to the Spanish Law) and on the vehicles/equipments for the carriage of perishable foodstuffs are available: both of them are not relevant for PESCaDO.

Regarding travel means, some relevant concepts are described in the following ontology, under the sub-hierarchy rooted at TransportationDevices:

DAML Transportation

Developed by: Teknowledge Corporation
Url: http://reliant.teknowledge.com/DAML/Transportation.owl

This ontology represents the transportation-related information contained in the CIA World Fact Book[6] (updated at the 2002 version). It was originally written in DAML, although an OWL version is available.

At the time of writing this document, the latest available release of this ontology is an unnumbered version (dated 19/08/2003), which includes:

- 444 concepts;
- 89 object properties;
- 4 data-type properties;
- 181 individuals.

The DL expressivity of the whole ontology is *ALCH(D)*. The ontology (at least the version tested) did not pass successfully the consistency check.

Here below we report a short extract of the taxonomy of transportation related entities:

Figure 8 - Excerpt of transportation related concepts in DAML Transportation

Although the whole ontology did not pass the consistency check, part of it may be reused for the purpose of PESCaDO.

Concerning weather-related road conditions, road traffic situation, road situation with respect to on-going roadworks, no already available ontologies have been found. Examples of the kind of information expected in the ontology regarding weather-related road conditions may be found here: http://alk.tiehallinto.fi/alk/english/frames/tiesaa-frame.html.

[4] José Javier Samper Zapater, Eduardo Carrillo Zambrano, Arturo Cervera García – "Semantic Modelling of Road Traffic Information elements" – in Euro-American Conference on Telematics and Information Systems 2006 (EATIS 2006)
[5] http://robotica.uv.es/~cicyt/
[6] https://www.cia.gov/library/publications/the-world-factbook/

1.3.5 Human Health

Concerning human health, there are several ontologies available. The focus of them is quite different, depending on whether they are meant to target specialists or doctors, rather than laypersons. Since one of the purposes of PESCaDO is to provide decision support to end-users, we have focused our investigation on health ontologies using a terminology for laypersons, and among them we point out an ontology describing human diseases and symptoms according to the ICD-10 standard international health care classification.

ICD-10 Ontology

Developed by: Data & Knowledge Management Unit, Fondazione Bruno Kessler (FBK)
Url: https://dkm.fbk.eu/index.php/ICPC2_Ontology
The ICD-10 Ontology is a formalization in OWL-DL of the International Classification of Diseases - 10th edition, published by the World Health Organization (WHO) in 2004 (http://www.who.int/classifications/icd/en/).

At the time of writing this document, the latest available release of this ontology is version 1.0 (dated 17/03/2008), which includes:

- 14502 concepts;
- no object properties;
- 9 data-type properties;
- no individuals.

The DL expressivity of the whole ontology is *ALC*. The ontology passed successfully the consistency check.

Following closely the ICD-10 classification, the ontology classifies diseases and symptoms in macro-categories (called *blocks*), depending on the part of the body related to the symptoms. Hence, for example, concept J00_J99 corresponds to *Diseases of the respiratory system*, or concept R00_R99 corresponds to *Symptoms, signs and abnormal clinical and laboratory findings*. A detailed description of the first hierarchy level of the taxonomy of concepts of the ICD-10 Ontology is presented in the table below:

Table 1 - ICD-10 Ontology top-level concepts and descriptions

Concept	Label
A00_B99	Certain infectious and parasitic diseases
C00_D48	Neoplasms
D50_D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
E00_E90	Endocrine, nutritional and metabolic diseases
F00_F99	Mental and behavioural disorders
G00_G99	Diseases of the nervous system
H00_H59	Diseases of the eye and adnexa
H60_H95	Diseases of the ear and mastoid process
I00_I99	Diseases of the circulatory system
J00_J99	Diseases of the respiratory system
K00_K93	Diseases of the digestive system
L00_L99	Diseases of the skin and subcutaneous tissue
M00_M99	Diseases of the musculoskeletal system and connective tissue

N00_N99	Diseases of the genitourinary system
O00_O99	Pregnancy, childbirth and the puerperium
P00_P96	Certain conditions originating in the perinatal period
Q00_Q99	Congenital malformations, deformations and chromosomal abnormalities
R00_R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
S00_T98	Injury, poisoning and certain other consequences of external causes
V01_Y98	External causes of morbidity and mortality
Z00_Z99	Factors influencing health status and contact with health services
U00_U99	Codes for special purposes

1.3.6 Geospatial Information and geographic datasets

Regarding geospatial information and geographic datasets, one of the most popular one is GeoNames, a large database for which an ontological schema and semantic web services are available.

GeoNames geographical database and ontology

Founded by: Marc Wick (founder)
Url: http://www.geonames.org/

The GeoNames geographical database contains over eight million geographical names from all over the world (more than 44.000 names are contained about Finland). The GeoNames toponyms have a unique URL with a corresponding RDF web service that allows accessing the geospatial semantic information attached to it. The kinds of geospatial semantic information attached to toponyms are defined in the GeoNames Ontology.

At the time of writing this document, the latest available release of the GeoNames Ontology is version 2.1 (dated 10/05/2010), which includes:

- 10 concepts;
- 13 object properties;
- 9 data-type properties;
- 681 individuals.

The DL expressivity of the whole ontology is ALH+(D). The ontology passed successfully the consistency check.

Typical examples of semantic information attached to toponyms are children/parent toponyms (e.g. the administrative subdivisions for a country) and nearby toponyms (e.g. names of locations close to the current one). An RDF dump (with 6.520.110 features and 93.896.732 triples)[7] of the semantic data contained in the GeoNames database can be downloaded from the GeoNames website: the dump has one RDF document per toponym on every line of the file.

[7] data collected on the 8th of June 2010

1.3.7 Other available environmental-related ontologies

The Environment Ontology

Developed by: The Environment Ontology Consortium
Url: http://www.environmentontology.org/

The Environment Ontology is a community-based ontology for describing the environment (habitat) of any organism or biological sample. It is particularly focused on the description of *biome*, i.e. a type of complex ecological community characterized by specific environmental conditions and a distinctive group of living beings, like e.g. the tundra biome of rain forest biome.

The latest available release at the time of writing this document is dated May 2009, and includes:

- 1239 concepts;
- 5 object properties;
- no data-type properties;
- 7075 individuals.

It is written according to the OBO file format (an alternative ontology representation language), but a derived OWL version is also available on the ontology web-site. The DL expressivity of the ontology is *ALE+*. The ontology passed successfully the consistency check.

Since the focus of this ontology is mainly on describing habitats and biomes, it may not be particularly relevant for PESCaDO, at least for the use cases considered.

1.4 Final remarks on environmental ontologies

As shown by this survey and first analysis on the state-of-the-art of environmental ontologies, several candidate ontologies are already available and could be used as part of the ontologies needed in PESCaDO.

In order to include them, in most of the cases, some (minor) adaptations may be required to tailor the ontologies for the specific PESCaDO purposes and use cases. For example, in the case of SWEET ontology, PESCaDO not relevant modules may be ruled out for efficiency reasons, while the missing information on some important pollutants will have to be added. Similarly, a rearrangement of the taxonomy of concepts may be required to adapt the ontologies to the domain considered, like e.g. in SWEET ontology it may be useful to collect all air pollutants under a common "AirPollutant" concept, instead of having them organized under chemical substances.

The meteorological phenomena and condition domain is comprehensively covered by SWEET ontology; nonetheless, we may considering integrating in PESCaDO also the other two ontologies for that domain here described: in this case, since some concepts are defined in more than one ontology (e.g. concept "Wind" is defined both in SWEET and MetBroker), techniques for ontology matching will have to be applied in order to merge correctly the selected ontologies.

It has to be noted that some relevant PESCaDO domains are not covered by any currently available ontology, like for example the pollens domain. That means these ontologies will have to be built from scratch, by applying some standard ontology building strategies:

- plan an ontology modelling phase, in which a team of domain experts and ontology engineers collaborate in the definition of the missing ontologies, possibly supported by some collaborative ontology engineering tools, like e.g. MoKi[8] or Collaborative Protégé[9];
- apply ontology learning techniques and tools to automatically extract ontology concepts and relations from structured (database schemas, glossaries) and unstructured (text documents, web sites) resources;
- apply a combination of the two previous strategies.

Another important remark concerns the relations between the different domains considered in PESCaDO. For example, certain meteorological/air quality conditions may be risky or cause some symptoms in people suffering from some health disease. In this case, the PESCaDO application, in order to provide adequate user decision support, will have to take into account also this kind of information. To the best of our knowledge, no ontologies exist which formalize this kind of relations, like the ones between e.g. environmental conditions and human health. We believe that the main reason for this is that, although effects of environmental conditions on human health are documented and reported in some documents and directives[10], the information provided is often vague and underspecified, and a precise structured description is lacking. Hence, even in this case, some strategies like the ones previously mentioned will have to be planned in order to obtain a formal description of this kind of knowledge.

[8] http://moki.fbk.eu
[9] http://smi-protege.stanford.edu/collab-protege/
[10] see for example, the "WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - Global update 2005 - Summary of risk assessment" – available for download at: http://whqlibdoc.who.int/hq/2006/WHO_SDE_PHE_OEH_06.02_eng.pdf

2 Inventory of corpora from the environmental domain

2.1 General characteristics of the corpora

To our knowledge, corpora from environmental domain are not available in electronic format for natural language processing (NLP) purposes. For this reason, it is necessary to acquire them semi-automatically from the web or to gather them manually from different sources. The corpora needed for the PESCaDO project present the following characteristics:

- They have to concern the environmental domain, and more specifically the topics dealt with in the project such as pollution, weather conditions, pollen situation, etc. We refer in particular to the Pilot Use Cases, which will be documented in Deliverable D8.5.
- Since the service orchestration task is focused on the activity of the Finnish Meteorological Institute, which will provide most of the domain-specific information for the project, the corpora and resources have to include, if possible, multilingual material in the three official languages of FMI documents, namely in English, Finnish and Swedish.
- Since the corpora are needed to acquire the linguistic knowledge necessary to the service orchestration and delivery task and to develop appropriate NLP tools for the content analysis steps, they have to be possibly large and representative of the terminology used in the project. The amount of data is crucial also to the creation of statistical NLP tools.

Apart from corpora, multilingual dictionaries represent very useful resources in different NLP tasks. In the PESCaDO project, they can provide a backbone for the automatic alignment of corpora in several languages as well as a database for the extraction of domain-specific documents and for their indexing. For this reason, the environmental dictionary EnDic[11] will be used, which contains 90,000 search words in different languages, the definition of 2,000 terms in English, Finnish and Estonian and 9,000 meteorological terms in Finnish, Swedish and English (previously part of the Meteorological Dictionary MetDic created by FMI).

2.2 Corpora inventory

The corpora collected include multilingual documents in .doc or .pdf format, which will be converted into plain text to be processed, and html documents downloaded from relevant websites in the environmental domain. The main topics to be included in the corpora inventory are the same as for the environmental ontologies (see Section 1.2). More specifically, they have to cover meteorological conditions and phenomena, air quality, pollens, travel and traffic information, health issues and geospatial information related to Finland's area considered in the project.

The two corpora types (official documents or html text from websites) will be described in the following subsections.

2.2.1 Official documents by environmental institutions

The pdf and doc documents collected are reported in the following table:

Document title	Number	Languages	Issuing Institution

[11] http://mot.kielikone.fi/mot/endic/netmot.exe?UI=ened

	of tokens		
Air quality in the Helsinki Metropolitan Area 2007	3,300	Fin, Swe, En	YTV (Helsinki Metropolitan Area Council)
AQ background action plan	23,500	Fin	YTV
AQ report Helsinki year 2008	19,200	Fin	YTV
Helsinki Region Housing Report 2008	9,400	Fin (with abstract in Fin, Swe and En)	YTV
Air Quality in the Helsinki Metropolitan Area in 2008	14,800	Fin (with abstract in Fin, Swe and En)	YTV
YTV Air Quality Action Plan for the Period 2008-2016	9,400	Fin (with abstract in Fin, Swe and En)	YTV
Helsinki Metropolitan Area Business Report	38,300	Fin (with abstract in Fin, Swe and En)	YTV
Helsinki Metropolitan Area Climate Strategy to the Year 2030	15,000	Fin and En	YTV
Air Quality in the Helsinki Metropolitan Area	3,800	En	YTV
Particles in the air	1,800	Fin and Swe	Finland's Environmental Administration
Smoke signals	2,500	Fin and Swe	Finland's Environmental Administration
Progress in Helsinki Metropolitan Area Climate Work – December 2009	4,300	En, Swe	YTV
Progress in Climate Work in the Helsinki Metropolitan Area – December 2008	3,500	Fin, En, Swe	YTV
Progress in Helsinki Metropolitan Area climate work – June 2009	4,200	En, Swe	YTV
What do we breathe?	2,800	Fin, Swe	YTV/FMI/HELI/Ministry of Environment
Vaasa Region Air Quality in 2007	6,000	Fin	Vaasa City Environmental Department
Vaasa Region Air Quality in 2008	6,000	Fin	Vaasa City Environmental Department
A Bioindicator study on the effects of air pollution in the Vaasa region during 2006-2007	19,200	Swe (with summary in En and Fin)	Vaasa City Environmental Department

Report on the Vaasan area benzopyrene air content	2,100	Fin	Vaasa City Environmental Department
40 documents with Vaasa city air quality monthly report	300 each	Swe	Vaasa City Environmental Department
Traffic Management Services Strategy	2,100	Fin	Road Administration Authority
Thoughts on the Air	12,000	Fin	Respiratory association
Metropolitan Area Company Overview	2,700	Fin	YTV
Helsinki Metropolitan Area Housing Report 2001	10,000	Fin	YTV
Helsinki Metropolitan Area Housing Report 2002	11,000	Fin	YTV
Helsinki Metropolitan Area Housing Report 2004	13,700	Fin	YTV
Helsinki Metropolitan Area Housing Report 2006	18,500	Fin	YTV
Future Work	26,000	Fin	YTV
Wood Heating	870	Fin	Respiratory Association
Pollen Allergy	1,700	Fin	Allergy and Asthma Association
Metropolitan Area Company Report (slides)	700	Fin	YTV
Metropolitan Area Air Quality 2009	1,100	Fin	YTV

The above documents have been collected from websites of official institutions or made available by FMI. Even if the size of the corpus they build is not very large if compared to multilingual NLP corpora used for statistical processing (ex. Europarl [Koehn, 2005]), it meets the requirements mentioned in Section 2.1. In fact, it covers the main topics dealt with in PESCaDO, i.e. meteorological, environmental, traffic and health issues, ranging from traffic reports to pollen classification. This will allow us to investigate and acquire the domain-specific terminology for the project. Another relevant aspect is the presence of multilingual material. In some cases, the original Finnish document is translated into Swedish and English, enabling us to straightforwardly creating parallel corpora with standard NLP tools for sentence and word alignment (for example, GIZA++ [Och and Ney, 2003]). In other cases, such documents are available only in Finnish or Swedish, which will be exploited for the collection of monolingual terminology and the development of language-specific tools. In the next project steps, the documents will be converted into plain text and, if multilingual material is available, it will be used to build parallel subcorpora.

2.2.2 Html documents from relevant websites

In order to collect documents related to different types of weather conditions and other aspects that may change day after day, another corpus is being created by extracting data from domain-specific websites. Similar to the other corpus, we mainly focused on websites available in the three languages of the project, namely Finnish, Swedish and English.

The list of the relevant links has been made available by FMI. The general approach was to extract once all general information from the site, and then to acquire once a day the data from the links which are frequently updated, for example about weather forecast or air condition. To this purpose, the WebDownload toolkit ([Girardi, 2010]) developed at FBK has been employed, which crawls through given web pages following some parameters defined by the user, stores their contents and eventually cleans the html code into plain text. In particular, the *Webdown* module generates as output a website archive in .cgt format (Catalog of Gzipped Texts), while other settings are stored in a logfile. Then, the *Webparser* application converts the content of the .cgt archive into a structured XML file where the html metadata are preserved as xml tags.

The extracted websites include:

http://www.ymparisto.fi: Finland's environmental administration website in Finnish, Swedish and English. The complete site has been downloaded (82 MB).

http://www.vaasa.fi/: Vaasa city website. There are versions in different languages, including Finnish, Swedish, English, Russian and German, which however are not parallel (Finnish and Swedish versions are richer and contain more links). The complete site has been downloaded (8.4 MB).

http://www.hengitysliitto.fi/: Finnish pulmonary association website with Finnish, Swedish and English version, which however are not aligned. The complete site has been downloaded (25 MB).

http://www.allergia.com/: Finnish allergy and asthma federation, in English, Swedish and Finnish. The complete site has been downloaded (69 MB).

The websites that have been downloaded and are daily updated are:

http://www.ilmanlaatu.fi/: Finland's air quality portal in English, Swedish and Finnish.

http://www.hsy.fi: Helsinki Region Environmental Service Authority website in three languages, with information about air quality, climate and urban environment.

http://www.polleninfo.org: Portal of allergy information and national alerts for all European countries. In English and German.

http://aerobiologia.utu.fi: Website of the aerobiology unit of the University of Turku in Finnish, Swedish and English.

http://www.ivl.se: Website of the Swedish Environmental Research Institute with information about climate, air, transport and environment.

Similar to the corpus typology described in Section 2.2.1, also in this case the websites were chosen in order to meet the initial requirements of PESCaDO, i.e. the multilinguality of the extracted material, the corpus dimension for statistical NLP purposes and the domain-specificity.

2.3 Use of environmental corpora in PESCaDO

After having collected a good amount of documents, the environmental corpora will be first converted into plain text and then they will be analyzed in order to extract the domain-specific terminology. As a first step, the KX system [Pianta and Tonelli, 2010], developed at FBK, will be used in order create a list of the most relevant keywords in the English corpus. This list will

be used to check and possibly extend the coverage of the environmental ontologies developed in PESCaDO, see Section 1.3. In addition to this step, we will also experiment with ontology learning techniques, with the aim of enriching manually crafted ontologies with new concepts automatically derived from the PESCaDO corpus. We will also try and exploit the EnDic dictionary to automatically enrich the PESCaDO ontology with information about how concepts are expressed in English, Swedish and Finnish.

Another important function of the collected corpus will be to analyse the ways in which the information relevant for the project is expressed in existing environmental sites. Such analysis will allow for developing the information extraction techniques which are needed for answering the user requests addressed by the PESCaDO system.

References

[Girardi, 2010]: Christian Girardi: The Webdownload tool. FBK Technical report, 2010.

[Koehn, 2005]: Philipp Koehn: Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit 2005.

[Och and Ney, 2003]: Franz Josef Och and Hermann Ney: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, vol. 29, n. 1, pp. 19-51, March 2003.

[Pianta and Tonelli, 2010]: Emanuele Pianta and Sara Tonelli: KX: A Flexible system for Keyphrase eXtraction. To appear in Proceedings of SemEval, July 2010.