Analysis and Retrieval of Multimodal Environmental Information for Personalized Decision Support

Stefanos Vrochidis

Informatics and Telematics Institute
Centre for Research and Technology Hellas

CLEF, 19-22th September 2011, Amsterdam, Netherlands

Outline of the Talk

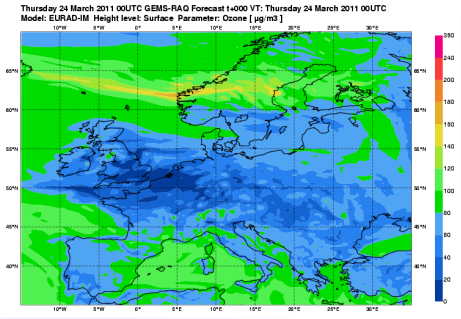
- Introduction
- PESCaDO project
- Use Case
- Discovery and Extraction of Environmental Information
- User Interaction Modes
- Conclusions

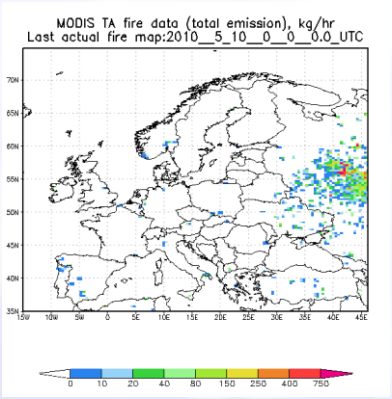
Introduction

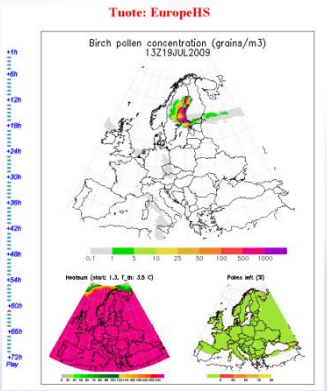
- Nowadays:
  - well-established national air quality and meteorological networks offer a variety of environmental services (i.e. air pollution, weather forecast, pollen).
  - Only a few of the data providers are giving access to their raw forecast and observation data.
  - A number of initiatives has been undertaken towards this direction, however the issue of harmonized access to environmental information remains open.
- The situation is not expected to change in the near future..
- The need today..
  - Citizens are increasingly aware of the influence of **air quality** and **meteorological conditions** on the **quality of their life** and demand for high quality environmental information that is tailored to one's specific context, background and needs.
  - E.g. People are highly interested in Pollen forecast.
    - More than 15% of the worldwide population suffer from pollen allergy.

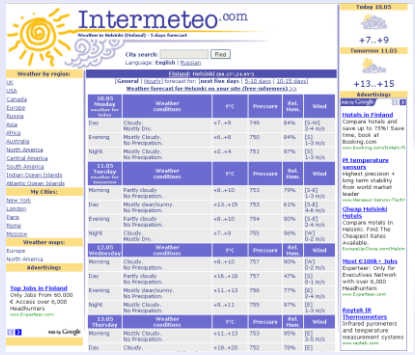
Environmental Nodes and Data

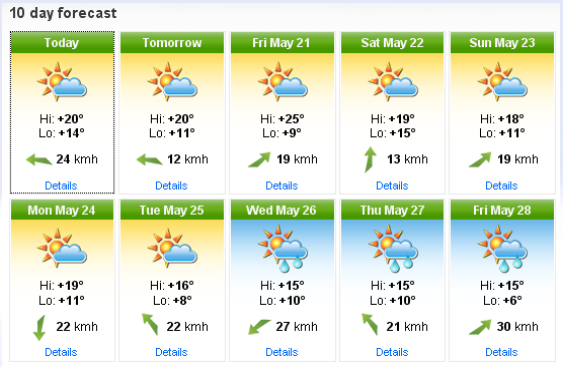
- Environmental Nodes
  - Websites with environmental information
    - Variety of encoding and representations
    - Several hundreds of websites
    - Usually free access
    - No detailed information of the covered area and stations employed
    - Multimodal Data
      - Free text
      - Text in tabular format
      - Images (heatmaps/graphs/icons)
      - Flash objects
  - Services with environmental information
    - Usually not freely accessible
    - They provide quality data
    - Detailed description of the stations is provided
    - Numerical structured data

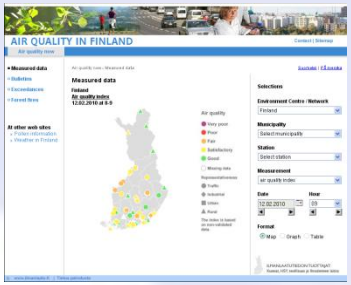






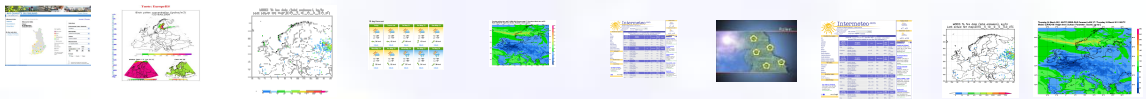






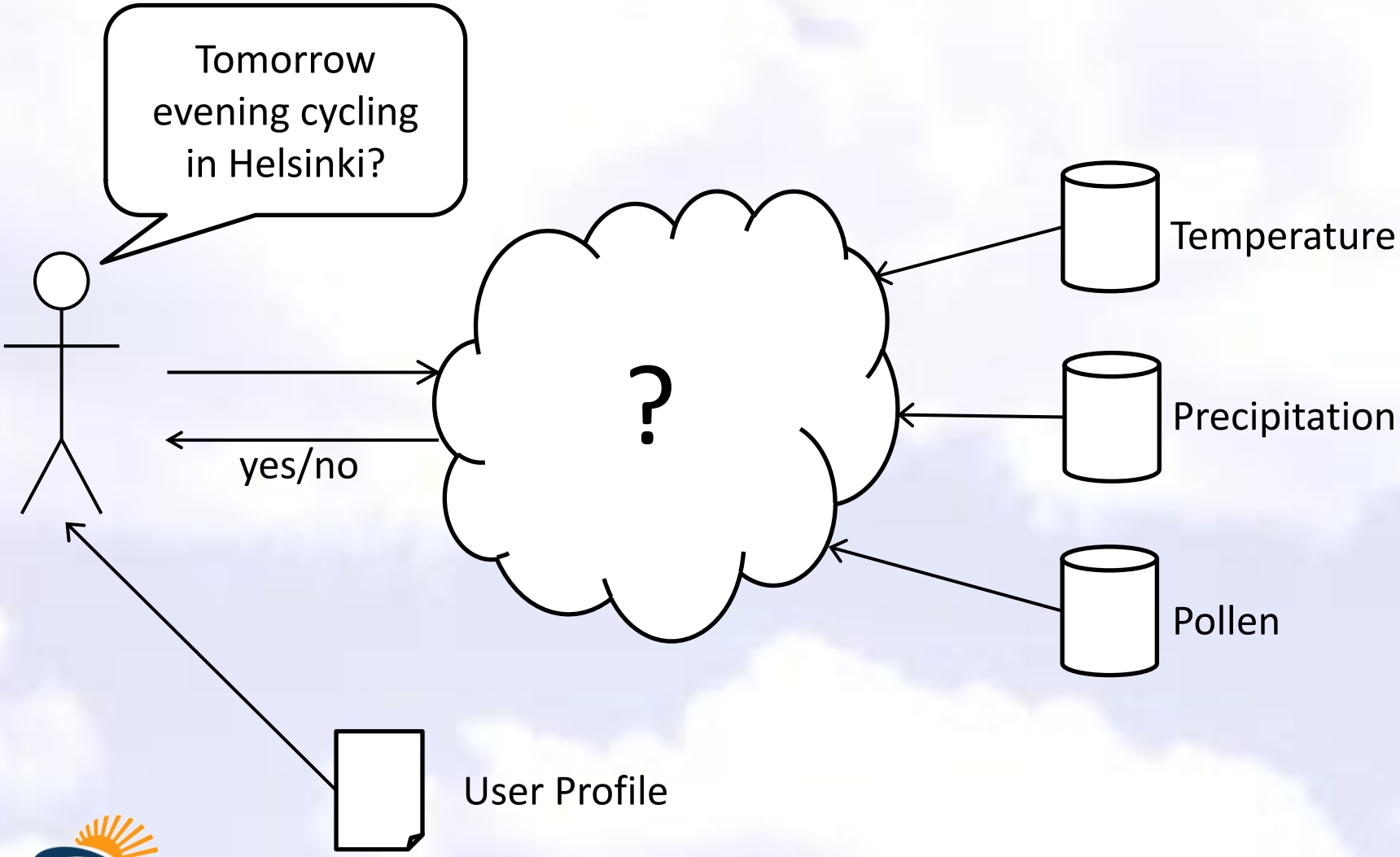
PESCaDO Project

- There is an increasing need for the orchestration of environmental nodes to provide users with personalized services supporting recommendations for environment and health conditioned tasks.

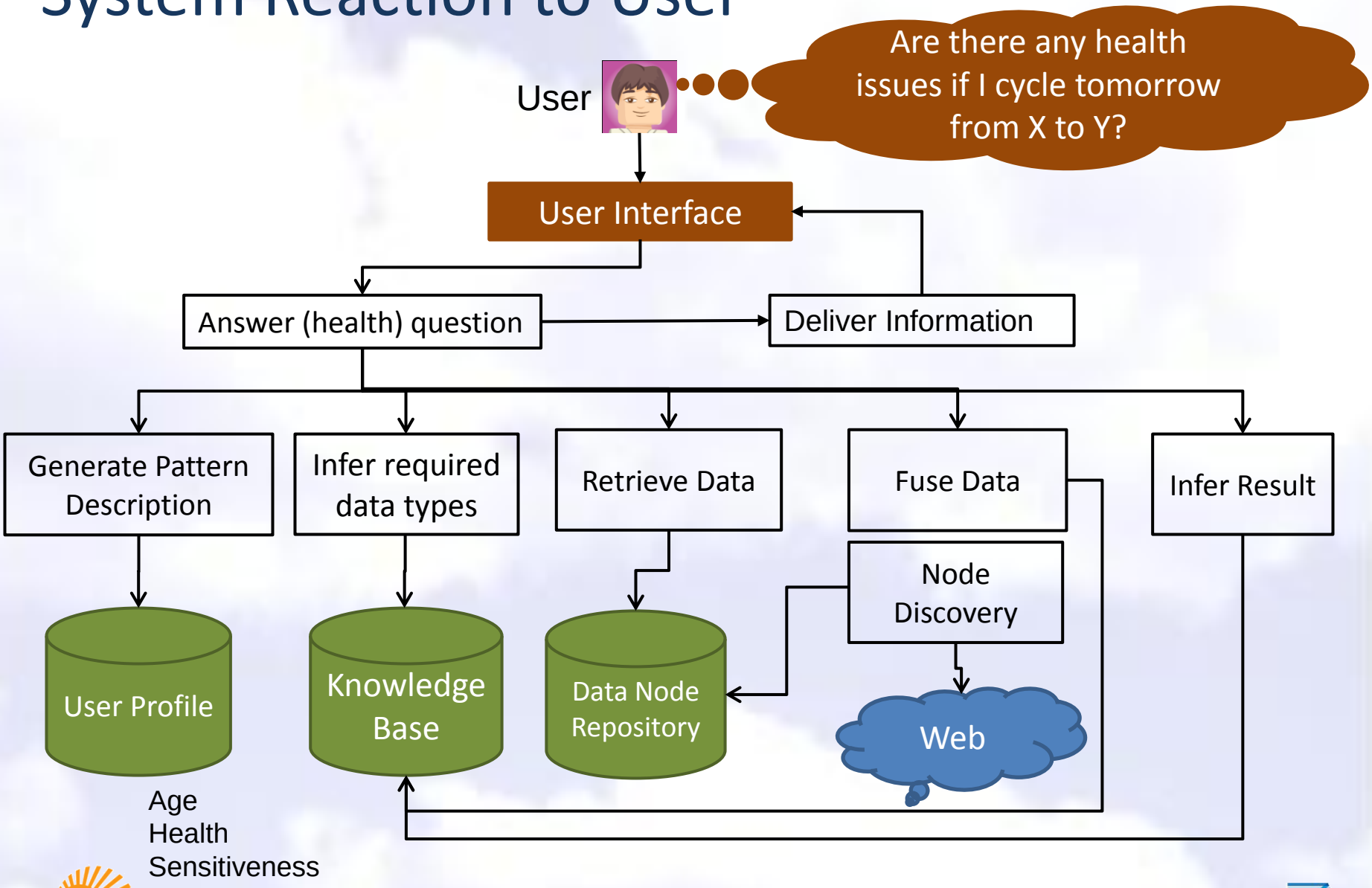


Personalized **E**nvironmental **S**ervice **C**onfiguration **a**nd **D**elivery **O**rchestration (PESCaDO) project addresses this challenge!

A Use Case …

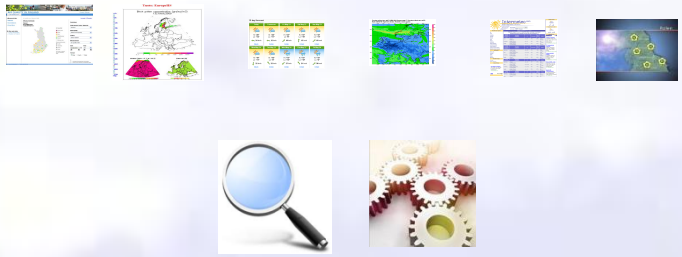

Tomorrow evening cycling in Helsinki?
?
yes/no
Temperature
Precipitation
Pollen
User Profile

System Reaction to User

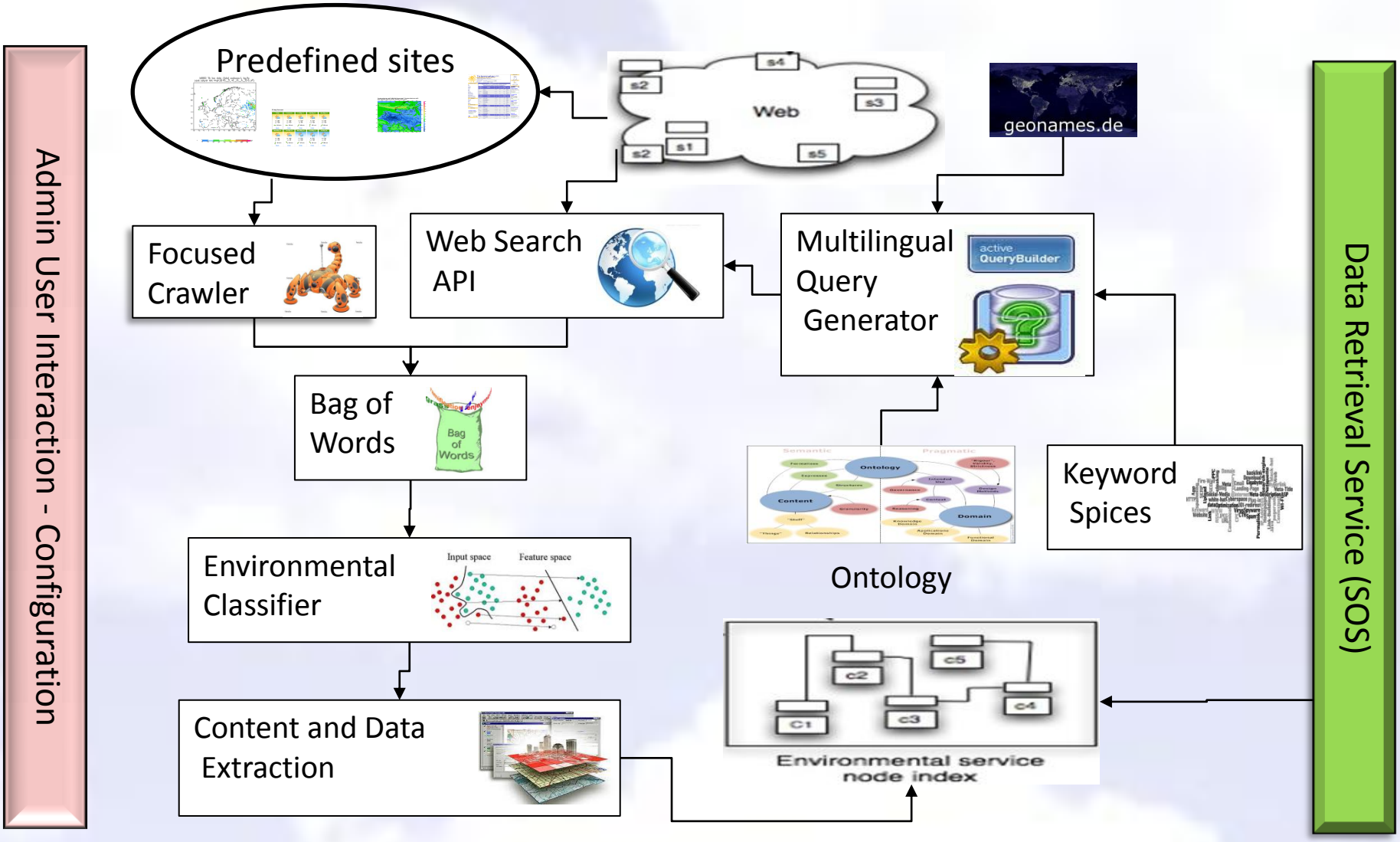


Environmental Nodes and Data

- A prerequisite for offering PESCaDO services are:
  - Discovery of Environmental Nodes
    - A problem of domain specific search
  - Content Extraction
    - Apply web text extraction and image analysis techniques



Environmental Node Discovery



Techniques for Domain Specific Search

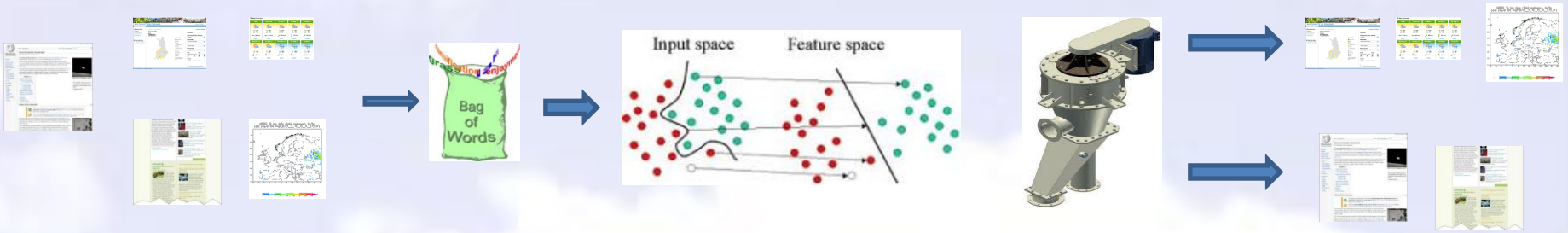
- Discovery of Environmental Nodes
  - Use a general purpose search engine

    - E.g. Yahoo, Google, etc.
    - Multilingual query generation using
      - Ontology
      - Keyword spice
      - Resources of geographical information (e.g. GeoNames, Foreca, etc.)
      - E.g. Helsinki + ozone + “air quality”
    - Application of post processing techniques
      - Classification
      - Filtering
  - Employ a focused crawler

    - Crawls the web in a directed way based on machine learning.
    - Based on a predefined set of websites (seed).
    - The crawler employs classification to collect only those Web pages belonging to a certain topic.

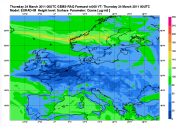
Techniques for Domain Specific Search

- Post processing based on Classification
  - Support Vector Machines are used for classification.
    - A model is trained using relevant and non relevant websites.
    - Classification can be employed also at lower level (e.g. weather classifier)
- Bag of words technique
  - Textual features generation using key-phrase extraction tools (KX)
    - Tokenization and N-grams extraction
    - Filtering of N-grams to select multiword expressions using morphological analysis
    - Ranking key-concepts based on normalized frequency and relevance



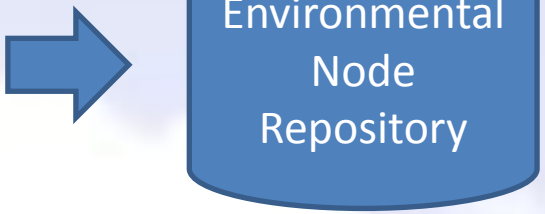
Empirical study of Environmental Nodes

- Environmental Website Statistics
  - Weather information: 80% of all information are reported in textual format or tables.
  - Air quality: 70% of the information about air quality is conveyed through images.
  - Pollen: 80% of the information about pollen is conveyed by images and maps.
  - No textual information at all for pollen and air quality forecast!

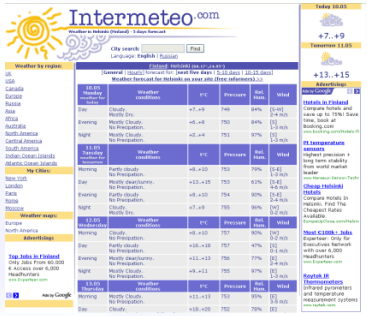


- Node Types
  - **Weather forecast**: temperature, pressure, wind, relative humidity
  - **Air Quality**: Pollutants: Sulphor Oxides (e.g. SO_2), Nitrogen Oxides (NO_2), Carbon Monoxide (CO), Dioxide (CO_2), etc.
  - **Pollen**: Grass pollen, birch pollen, alder pollen, etc.
  - **Alerts** (extreme weather conditions or phenomena): Extreme temperature, avalanches, coastal events, thunderstorms and snow/ice.

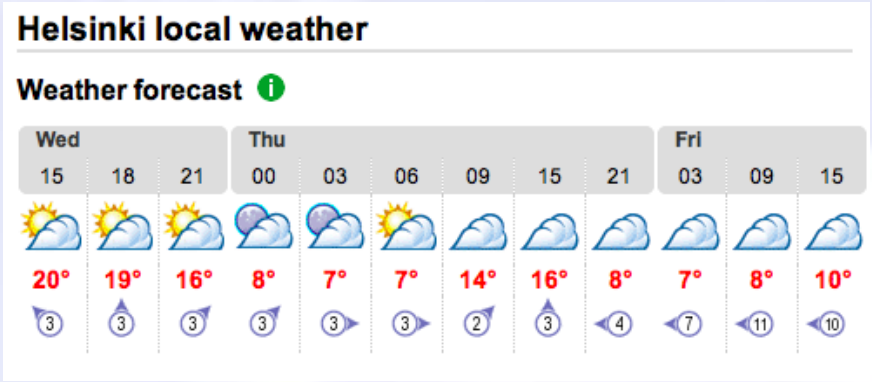
- The important information to be extracted
  - Type of measurement (e.g. temperature)
  - Type of data (historical, observed, statistical, forecast)
  - Measurement and Unit (e.g. 19^{o}C)
  - Geographic Location of measurement (e.g. Helsinki)
  - Date/time (e.g. 14:00, 3/9/2011)



Techniques for Textual Data Extraction



- Processing webpage content
  - Process content and metadata
  - Parsing the HTML structure of the retrieved pages
  - Extraction of the relevant information from either the textual content or the HTML structure.
    - Filtering out irrelevant sections (menus, boilerplate, advertisement, etc.)
    - Regular expressions
    - Tree representation to capture the webpage structure
    - Combination



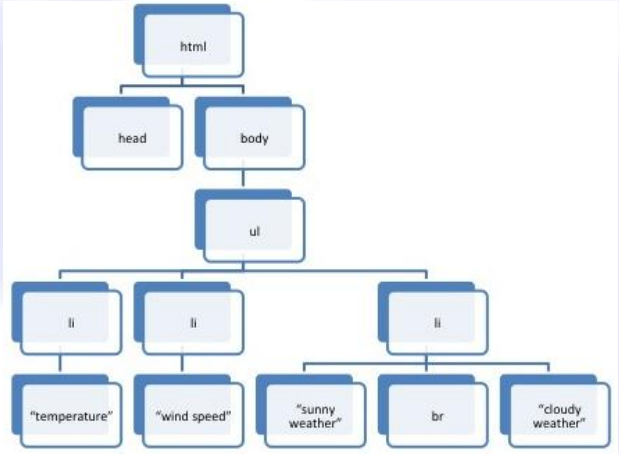
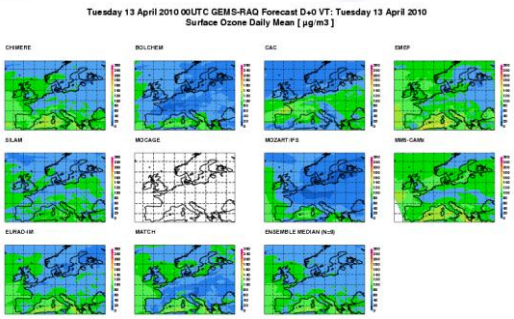
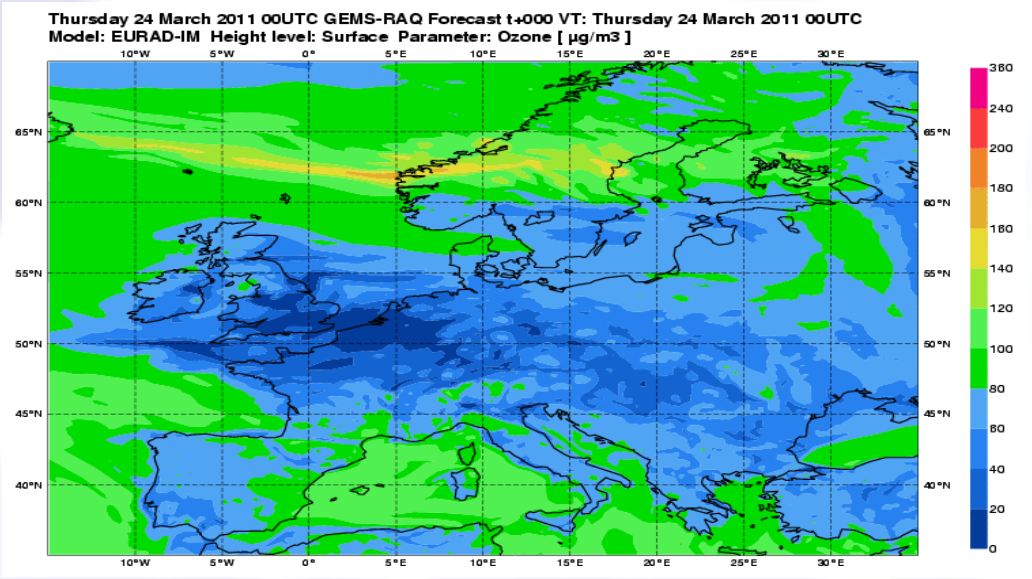


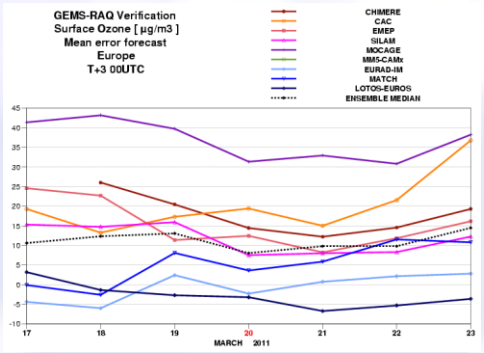
Image Types

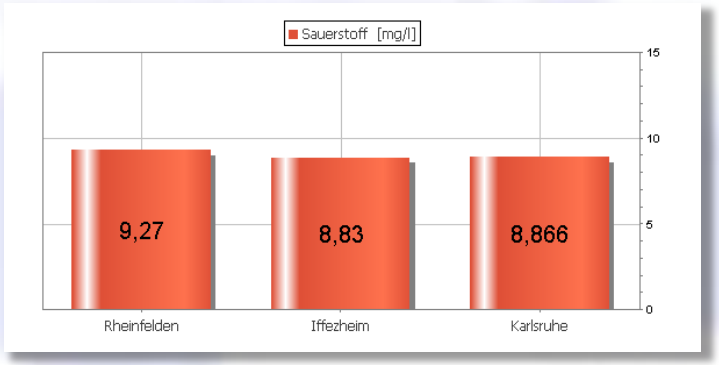
- Heat Maps





- Variety of Diagrams





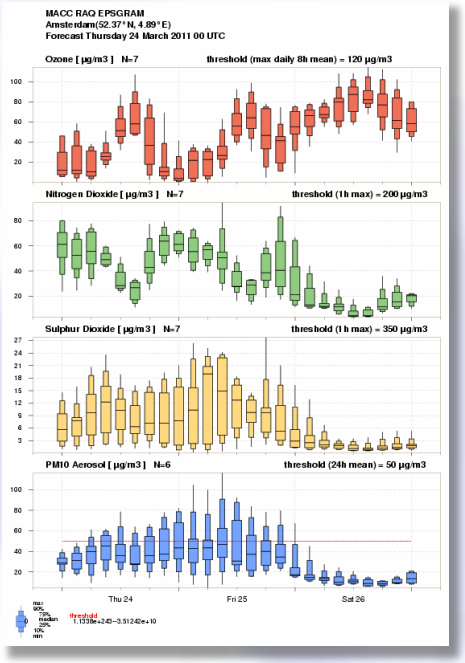
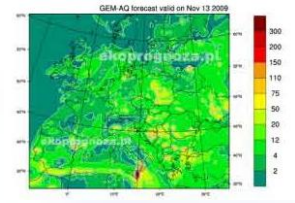
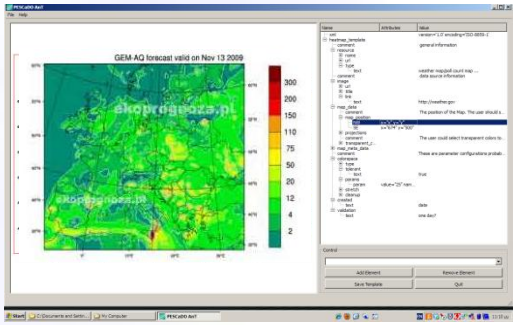


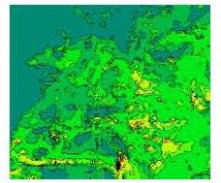
Image Content Distillation

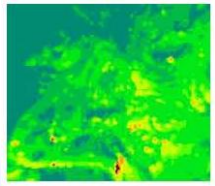
- The large variety of images leads us to a user-assisted approach.
- We adopt a template-based representation (XML) for each image category (e.g. heatmap, line graph).
- We employ an annotation tool for template configuration
  - User assisted segmentation
  - User assisted geopinning
- OCR-based service to identify axis information, title, time and date.
- Mapping the index values to specific coordinates.
- Reconstruction of missing values and data gaps using interpolation.





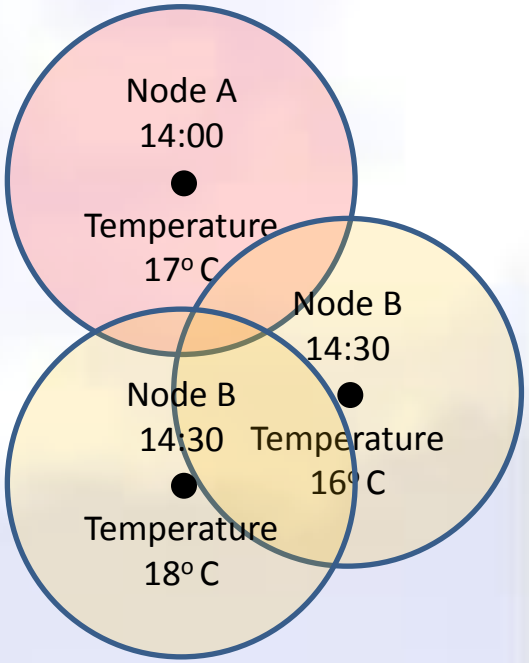
XML
template



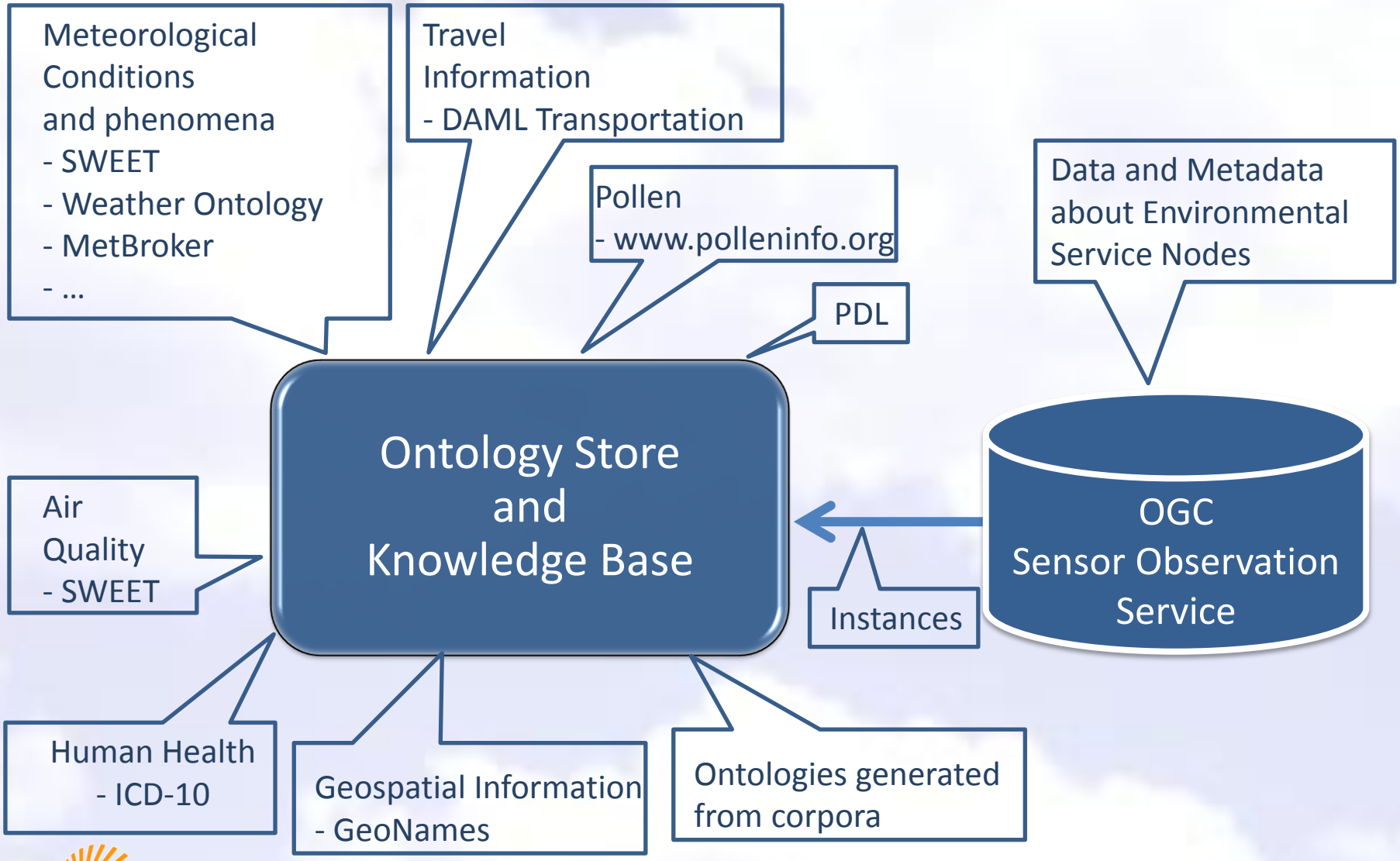


Environmental Data Fusion

- Fusion of environmental data is very complicated task
  - Environmental service nodes may provide on the same aspect for the same or the neighboring location:
    - Competing data
    - Complementary data
    - Different types of measurements (averages or exact)
    - For different time ranges
  - To ensure the availability of a most reliable and comprehensive content the content proceeding from these nodes must be fused.
  - This implies an assessment of the quality of the contributing services and data.
  - Quality metrics will be based on the evaluation of environmental nodes in time
    - Comparison of forecast data with observed data
    - Comparison of the fluctuation of a certain forecast over time

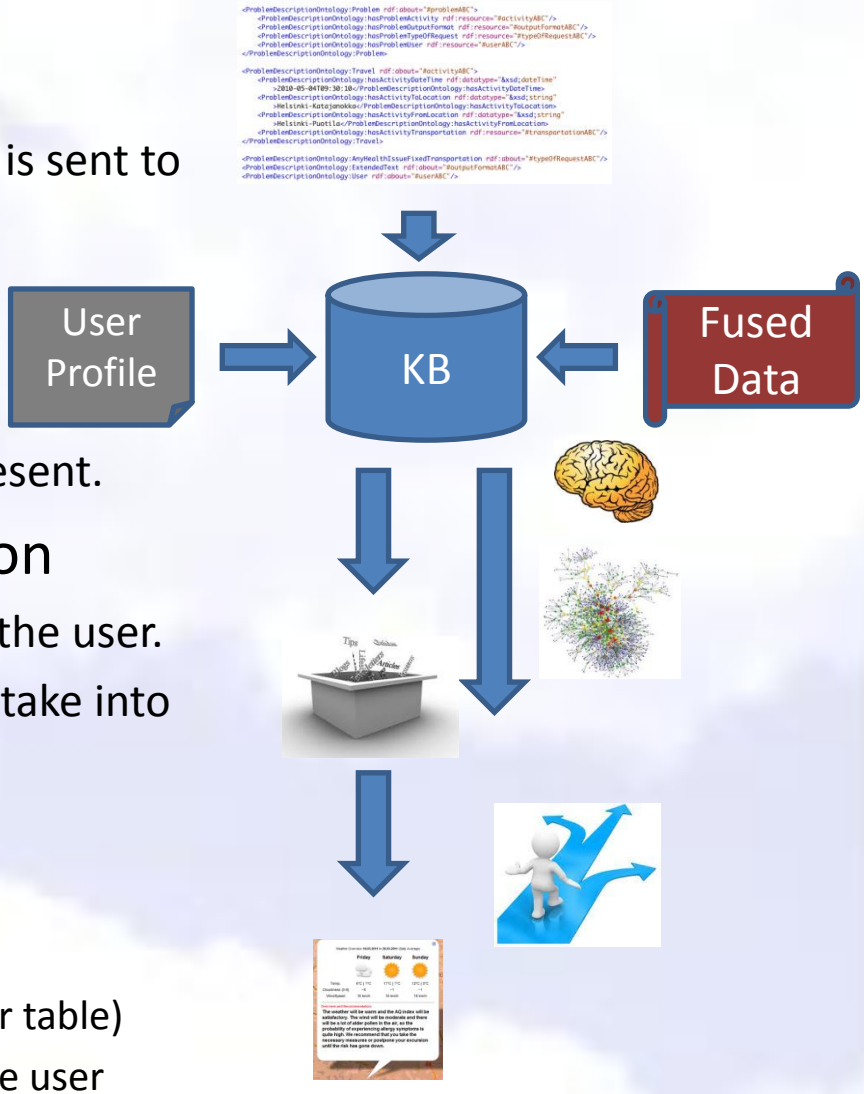


Knowledge Base

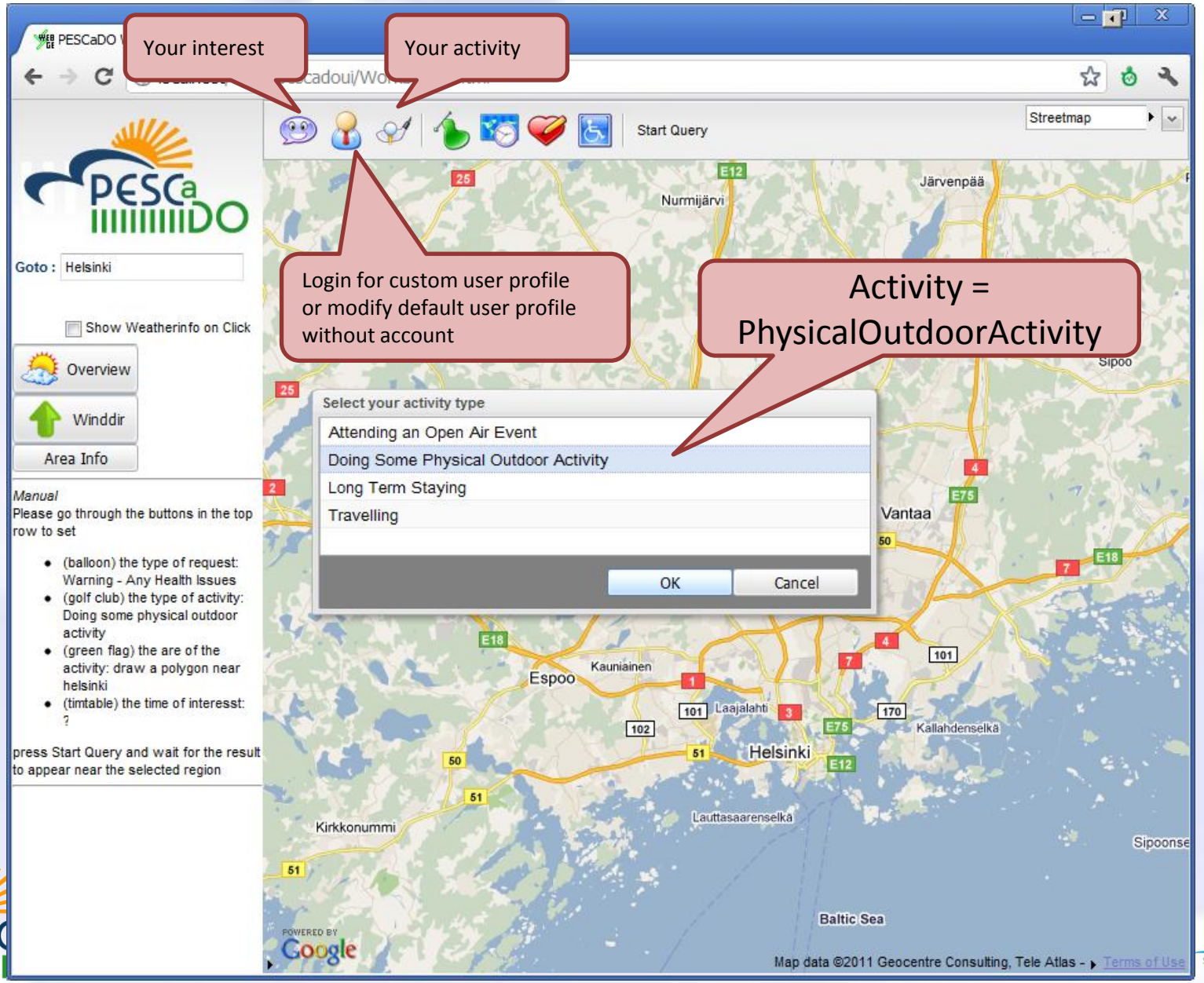


Decision Support and Content Delivery

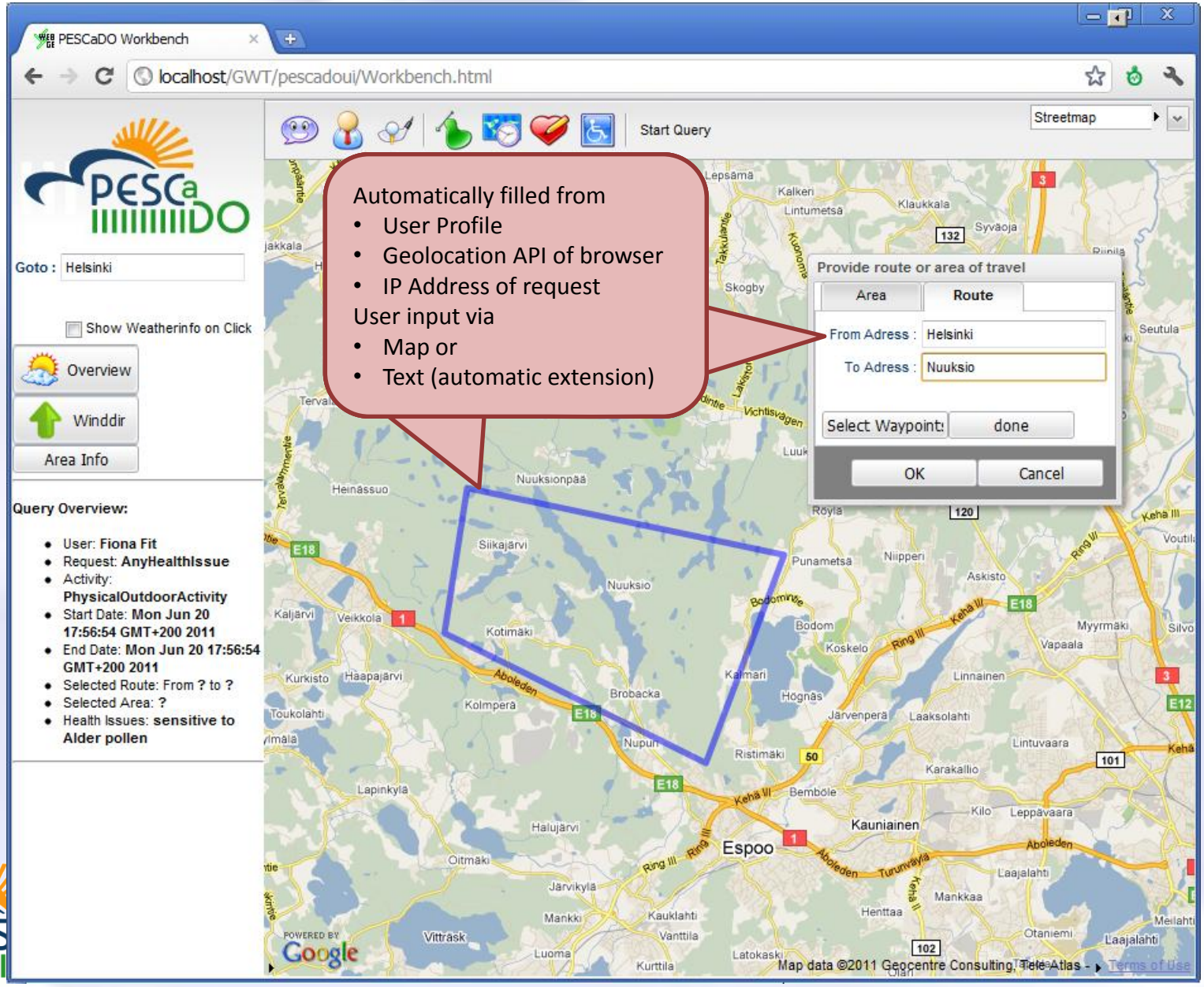
- Decision Support
  - An RDF Problem Description Language is sent to the KB.
  - The environmental data and the user profile are loaded on the KB.
  - Reasoning is performed in the KB after all the required information is present.
- Automatic Information Production
  - Not all content in the KB is relevant to the user.
  - Intelligent content selection strategies take into account
    - the background of the user
    - the intended use of the information
  - The content is delivered to the user:
    - in a suitable mode (text, graphic and/or table)
    - in the language of the preference of the user



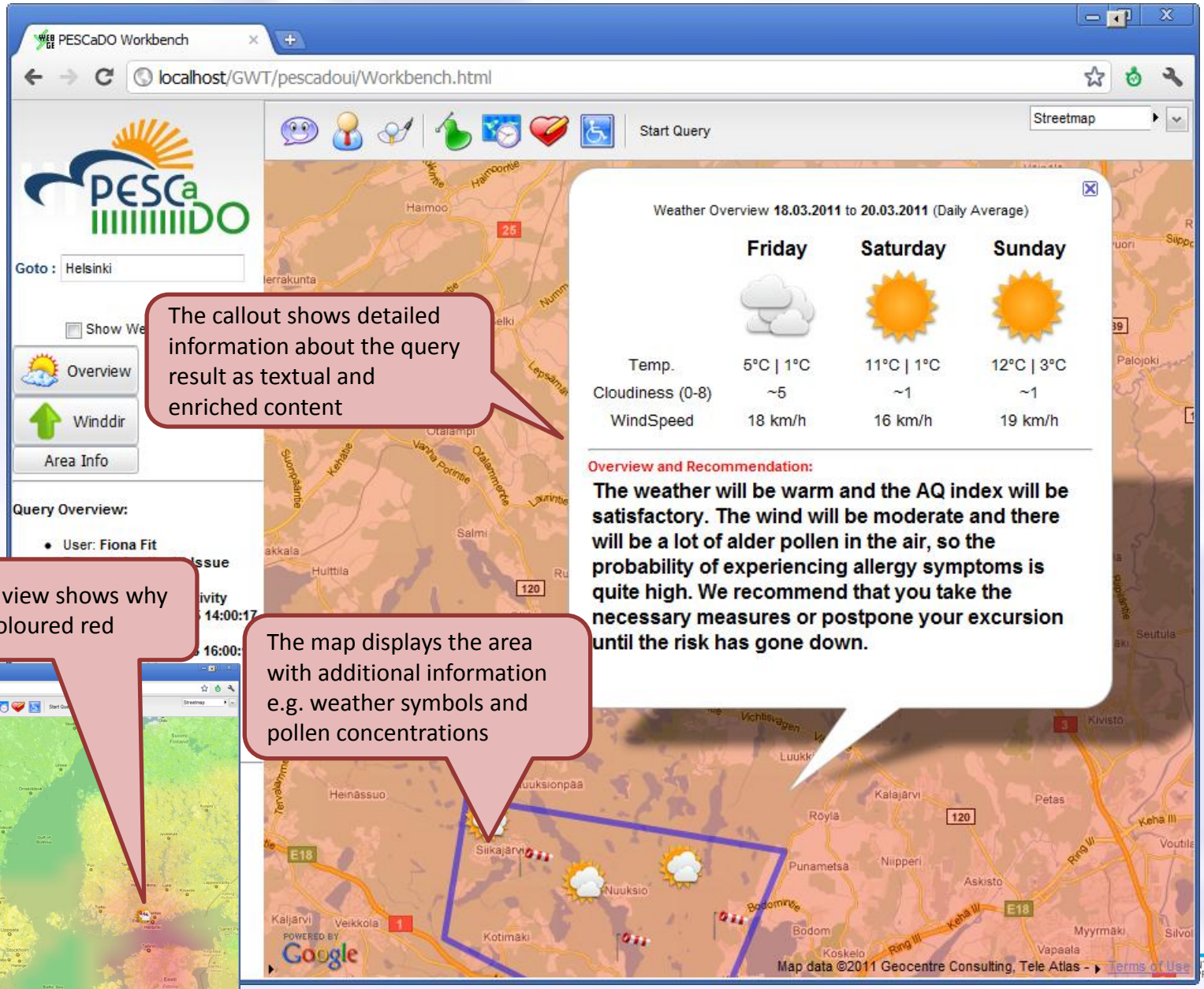
User Interaction – Select Activity



User Interaction – Build Question



User Interaction – Result Presentation



Conclusions

- People are increasingly interested in environmental information and therefore environmental applications are of great interest to the public.
- Some of the important challenges are:
  - The discovery of environmental nodes
  - The extraction of multimodal environmental information
  - The fusion of environmental information
- The variety of representations and lack of standards in the presentation of environmental information make these tasks very complicated.
- Several techniques from text mining , image analysis and information retrieval areas could be applied in the environmental domain
  - These techniques need to be optimized to deal with environmental data.
  - They could be of added value for the development of environmental related services

- **Interdisciplinary** research in the **Environmental** domain and **Information Retrieval** area could be beneficial for humanity in the context of offering services and solutions for **healthcare** issues related to **environmental** conditions.

PESCaDO Partners

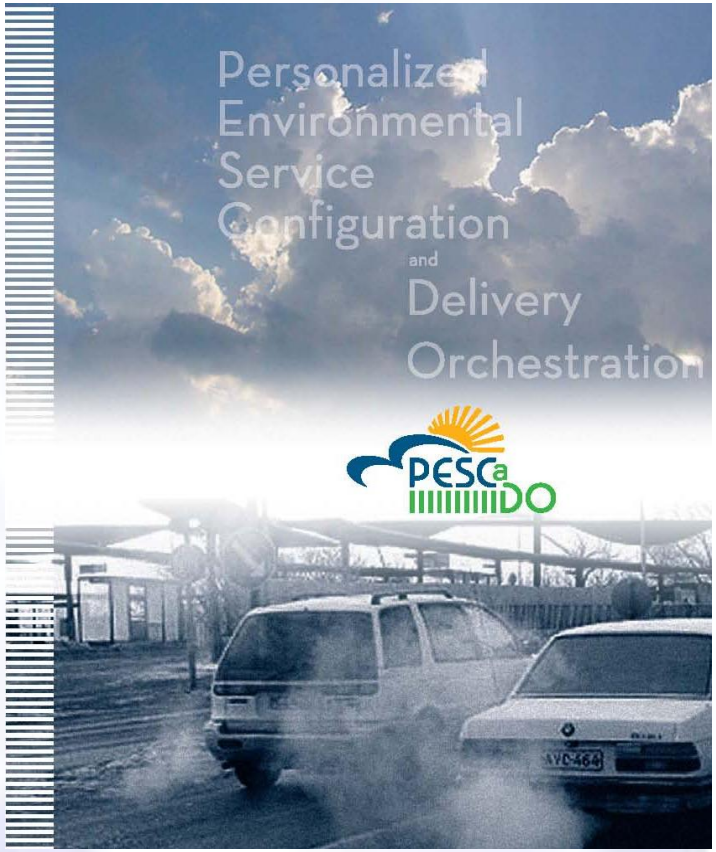
Universitat Pompeu Fabra (ES), (Coordinator)	UPF
Fraunhofer IOSB (D)	IOSB
Finnish Meteorological Institute (FIN)	FMI
Institute for Visualization and Interactive Systems, University of Stuttgart (D)	USTUTT
Fondazione Bruno Kessler (IT)	FBK
Centre for Research and Technology Hellas, Informatics and Telematics Institute (GR)	CERTH
Helsinki Region Environmental Services Authority (FIN)	HSY

External Cooperations

Informatics Systems and Application Group, AUTH (GR)	ISAG

Contacts

Leo Wanner (Coordinator), UPF, leo.wanner@upf.edu
Stefanos Vrochidis, ITI-CERTH, stefanos@iti.gr
Ioannis Kompatsiaris, ITI-CERTH, ikom@iti.gr



THANK YOU FOR YOUR ATTENTION

Questions?

http://www.pescado-project.eu