



D3.2 Report on required and achievable data quality and monitoring quality indicators

D3.2 covers: D3.2a and D3.2b

Work package	WP3		
Task	T3.2 and T3.3		
Dissemination level	<input checked="" type="checkbox"/> Public	<input type="checkbox"/> Restricted to programme	
	<input type="checkbox"/> Restricted to specific group	<input type="checkbox"/> Confidential	
Publishing date	Contractual: 01-03-2009		Actual: 28-02-2009
Deliverable	D3.2	Version 1	Draft <input type="checkbox"/> Final <input checked="" type="checkbox"/>
WP / Task responsible	Ramiro Neves		
Contact person	Cláudia Neto		
Contributors	Cláudia Neto (IST), Rodrigo Fernandes (IST), Susana Nunes (IST), Constança Belchior (IST), Francesco Archetti (UMB), Luigi Quarenghi (UMB), Sander Loos (HL).		
Short abstract	This document describes the principles related to data quality for the data manipulated in lenvis, in particular environmental data. For water and air data are described the collection and quality management procedures that are actually applied in Europe. A short overview of data quality requirement expressed by lenvis users is also given.		
Keywords	Data quality, data monitoring, data quality indicators		
Document	D3 2-Report-V01-reviewed-IST.doc		

Project Coordinator
 HydroLogic BV
 P.O.Box 2177
 3800 CD Amersfoort
 The Netherlands
 T: +31 33 4753535
 www.hydrologic.com



Table of contents

D3.2 Report on required and achievable data quality and monitoring quality indicators	i
D3.2 covers: D3.2a and D3.2b.....	i
1. Introduction.....	1
1.1. Principles relating to data quality	1
1.2. Difficulties in data modelling and data quality assurance	1
2. Data quality: the lenvis conceptual model	2
3. Quality of air data	4
3.1. Air pollution monitoring networks	4
3.2. Methods to estimate data source quality.....	7
3.3. Choice of reference values for air pollution	7
3.4. Quality issues in air pollution data	7
4. Quality of water data.....	10
4.1. Monitoring data	11
4.1.1. Classic sampling.....	11
4.1.2. Automatic stations.....	14
4.1.3. Profiles using sensors	14
4.2. Reference values in water quality.....	15
5. Data quality management	17
5.1. Data quality management in air pollution monitoring networks	17
5.2. Data quality management in wireless sensor networks	20
5.3. Data quality management in water data.....	21
5.3.1. Data quality management to classic sampling results	21
5.3.2. Data quality management in automatic stations data	22
5.3.3. Data quality management in profiles.....	23
5.3.4. MetaDatabase	24
5.4. Water quantity data management	26
6. Data quality requirements from users	27
7. References.....	28

1. Introduction

1.1. Principles relating to data quality

The next citation is from the Article 6 of the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995. It is the reference text on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Some quality requirements are already laid down in the same directive that sets the privacy guidelines.

Member States shall provide that personal data must be: (a) processed fairly and lawfully; (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards; (c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed; (d) accurate and, where necessary, kept up to date; every reasonable measure must be taken to ensure that data inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified; (e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use. It shall be responsibility of controller to ensure that paragraph 1 is complied with.

1.2. Difficulties in data modelling and data quality assurance

The issues of data quality and their assurance take different forms in different architectures and according to different types of data. As we see in the picture below (Fig. 1), the data quality requirements of different applications move along three dimensions:

- Distribution of the data, measured by the number of sources;
- Autonomy of the different data sources (vs. integration);
- Heterogeneity: a measure of the number of data types that have to be manipulated.

In the case of lenvis we have to face with the most difficult data quality issues, since we have to deal with many distributed data sources, most of them created and maintained autonomously, containing different types of data that are shared also through a peer to peer network.

Regarding data quality from database perspective, we must consider two aspects:

- Quality of data: typing errors, missing values, attribute exchange among records.
- Quality of the schema: it is important to structure the data following a rigorous model approach (normal forms).

An important property of data is the accuracy i.e. the correctness of the representation of the real-life phenomenon in the database. It refers to two different dimensions:

- Syntactic accuracy: check if every value is a permissible value in the corresponding definition domain.
- Semantic accuracy: closeness of the actual value of an attribute to the true value.

Another important property of data is the completeness (which addresses the problem of missing values). Another relevant aspect of the data is their frequency of change and update; three time-related dimensions characterize this aspect:

- Currency: how promptly data are updated;

$$\text{Currency} = \text{Age} + (\text{DeliveryTime} - \text{InputTime})$$

- Volatility: frequency with which data vary in time;

$$\text{Volatility} = \text{length of time that data remain valid}$$

- Timeliness: how data are useful for a given task. It is related to the concept of *real time*, i.e. the data have to be available when (or before) the user's application needs them.

$$\text{Timeliness} = \max\left(0.1 - \frac{\text{currency}}{\text{volatility}}\right)$$

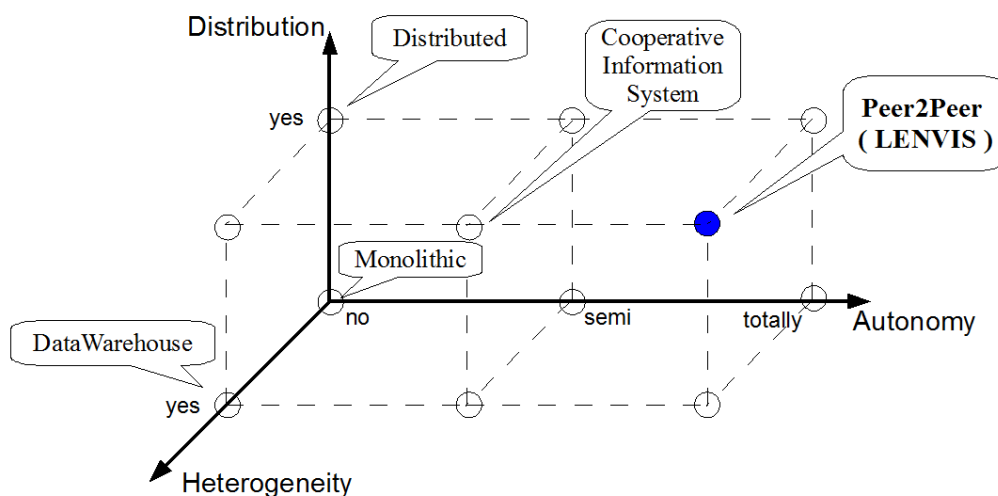


Fig. 1. Dimensions along with it is possible to measure the data quality requirements of different applications.

2. Data quality: the lenvis conceptual model

Data quality issues in lenvis arise from three areas:

- 1) **Air quality data:** these data are captured using a number of sensors. The two main reference scenarios are:
 - i. **Pollution monitoring networks:** a relatively small number of “large”, reliable environment “stations”, each equipped with a number of sensors of typically good accuracy, transmitting all the data to a centralized database (*sink*).
 - ii. **Sensor networks:** typically Wireless Sensor Networks (WSN), composed by a relatively large number of cheap devices. In this case data quality is affected by sensor failures or energy shortage as well as by the communication protocol errors and the data management faults both at the local (device) and global (network) level. The global level represents the sink where queries are formulated and replies are produced.

Data management protocols as well as the application design change according to the number of devices, the local computational power and the available network bandwidth.

In the first case the data are sent to a data warehouse after a limited pre-processing. In the second case a local model of the environmental data can be computed locally and used for forecasting and anomaly detection.

- 2) **Water quality data:** Water quality data can be acquired by classic sampling, automatic stations and sampling using sensors (profiles along the water column, x, y, or time).
 - i. **Classic sampling** - discrete variability in time and in space;
 - ii. **Automatic stations** - continuous variability in time, discrete in space;
 - iii. **Profiles using sensors**- continuous variability in time and space.
- 3) **Health data:** data quality problems in health data arise from their highly composite nature, i.e. they come from many heterogeneous sources, a very long time span from the generation of the data and their utilization, different data collection methods and the relevance of human intervention, leading to incompleteness and missing data.

In both cases, for environment (Air and Water) and health, if the data are incomplete or incorrect, derived decisions are likely to be faulty. However, while the issue of data privacy is relevant for clinical information, assessing and enhancing the quality of environmental data is fundamental for lenvis.

There are two modes of data processing:

1. Data are “consumed” directly from the observer, for basic data analysis (streaming data analysis);
2. The data are aggregated in a persistent data base, where complex data mining and knowledge discovery tasks are performed.

For health data only mode 2) applies, while both operation modes can be considered for environmental data. Lenvis is concerned with complex knowledge discovery tasks, mainly to discover time stamped causal relations between environmental data and health data, as measured by the hospitalization records, and therefore its main focus is to work on persistent data stored in a data-warehouse. Still the need to account for the dynamic nature of data makes relevant the “data streaming view” both for the quality issues and for the analysis algorithms developed in WP6.

Therefore the conceptual model that will be developed in this project for data quality and management consists of 4 steps:

1. Data quality recording
2. Data quality propagation in streaming data
3. Extension to the DBMS to account for quality of persistent data

3. Quality of air data

3.1. Air pollution monitoring networks

The environmental stations for air pollution considered in the Italian case study of lenvis are provided by Project Automation. Their architectural structure is depicted in Fig. 2. The pollutants that are monitored are represented from Table 1 to Table 4.

In Italy the characteristics of the monitoring stations are established by law that defines through qualitative criteria four main types of stations:

- type A: urban background station;
- type B: high density population station;
- type C: high traffic station;
- type D: suburban photochemical.

Each city has a different number of stations (ranging from 7 in Bologna and Palermo to 19 in Genoa), with a different distribution of station types. Comparability among networks is overall poor, also given the lack of national standards that define station locations. In particular, type B and C stations, which are the most relevant for the present exercise, are generally not comparable because the two criteria to define them are not mutually exclusive. Moreover, the traffic conditions and local meteorology vary by site.

Measurements of PM10 have been introduced very recently, i.e. January 1998 for most of the cities. Before this date, mainly TSP (Total Suspended Particulate) data are available. For some cities, PM10 measurements are still not available, or they are available for very few monitors of the network. In addition, there are no national rules that define air quality data collection methods for both TSP and PM10. At least two different methods are used to measure TSP (gravimetric and beta-ray absorption) and three for PM10 (gravimetric, TEOM, beta ray absorption). This study attempted to provide a reasonable interpretation of the existing data, which will be updated as monitoring data become more consistent.

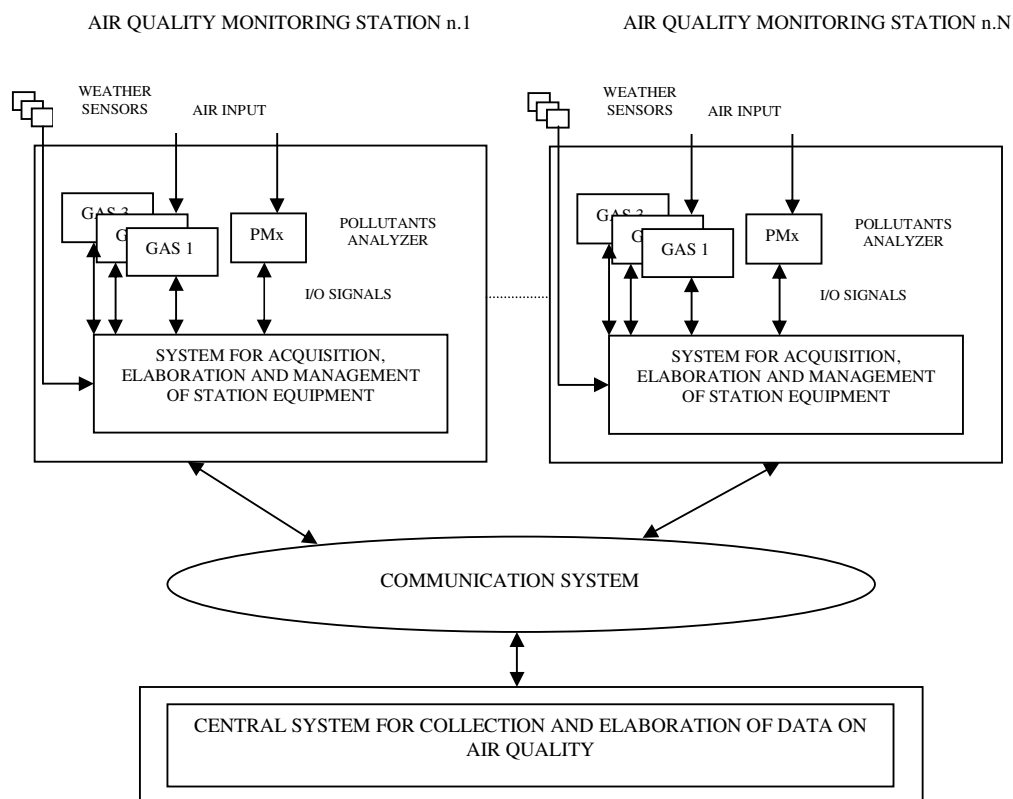


Fig. 2. Architectural organization of the environmental stations for air pollution.

Table 1. Organic air pollutants monitored.

Organic air pollutants			
<i>Pollutant</i>	<i>Automatically measurable</i>	<i>Measured in PA stations</i>	<i>Notes</i>
Acrylonitrile	NO	-	
Benzene	YES	YES	
Butadiene	YES	Few applications	There are no certified measurement instruments
Carbon disulfide	NO	-	
Carbon monoxide	YES	YES	
1,2-Dichloroethane	YES	NO	
Dichloromethane	YES	NO	
Formaldehyde	YES	NO	
Polycyclic aromatic hydrocarbons (PAHs)	YES	YES	There are no certified measurement instruments
Polychlorinated biphenyls (PCBs)	NO	-	
Polychlorinated dibenzodioxins and dibenzofurans (PCDDs/PCDFs)	NO	-	

Styrene	YES	Few applications	There are no certified measurement instruments
Tetrachloroethylene	NO	-	
Toluene	YES	YES	
Trichloroethylene	NO	-	
Vinyl chloride	NO	-	

Table 2. Inorganic air pollutants monitored.

Inorganic air pollutants			
<i>Pollutant</i>	<i>Automatically measurable</i>	<i>Measured in PA stations</i>	<i>Notes</i>
Arsenic	NO	-	
Asbestos 1	NO	-	
Cadmium	NO	-	
Chromium	NO	-	
Fluoride	NO	-	
Hydrogen sulfide 1	NO	-	
Lead	NO	-	
Manganese	NO	-	
Mercury	YES	NO	
Nickel	NO	-	
Platinum	NO	-	
Vanadium	NO	-	

Table 3. Classical air pollutants monitored.

Classical air pollutants			
<i>Pollutant</i>	<i>Automatically measurable</i>	<i>Measured in PA stations</i>	<i>Notes</i>
Nitrogen dioxide	YES	YES	
Ozone and other photochemical oxidants	YES	YES	
Particulate matter	YES	YES	
Sulfur dioxide	YES	YES	

Table 4. Indoor air pollutants monitored.

Indoor air pollutants			
<i>Pollutant</i>	<i>Automatically measurable</i>	<i>Measured in PA stations</i>	<i>Notes</i>
Environmental tobacco smoke	NO	-	
Man-made vitreous fibres	NO	-	
Radon	NO	-	

3.2. Methods to estimate data source quality

The functional parameters of these stations are:

- Quantitative measures
- Sampling rates
- MTBF (Mean Time Before Failure) / MTTR (Mean Time To Recovery)
- Error (variance)

The following general criteria were adopted for the selection of the stations:

1. The stations should be located within the city borders and very close to the population centroid.
2. The stations cannot be placed too close to local source emissions, particularly industrial, and they should be generally located in residential areas.
3. The daily correlation among the selected stations should be reasonably high ($\sim > 0.7$), to exclude outliers or monitors measuring “hot spots” instead of regional background concentrations.
4. The stations should provide a sufficient number of data ($\sim > 50\%$ of possible data must be valid for a given period of time).

To ensure an adequate representation of the population exposure, at least two stations are to be selected for each city.

3.3. Choice of reference values for air pollution

Three different environmental standards are used as reference levels for air pollution. The main reference level was chosen a priori at $30 \mu\text{g}/\text{m}^3$. This value is in between the European Union proposed Stage 1 and Stage 2 limit values of 40 and $20 \mu\text{g}/\text{m}^3$ PM10, respectively.

Using the environmental standards recently promulgated by the U.S. Environmental Protection Agency for PM2.5, the annual average standard for PM2.5 is $15 \mu\text{g}/\text{m}^3$. Assuming a ratio between PM2.5 and PM10 – approximately 0.5 (U.S. EPA, 1996) – this would be equivalent to $30 \mu\text{g}/\text{m}^3$ PM10. To estimate the attributable health effects from air pollution exposure, the change from these three presumed effect levels to the existing ambient concentrations in each city was calculated, based on current monitoring information.

It should be stressed, however, that all impact estimates calculated using these standards understate the overall health burden associated to exposure to PM10 in urban areas. As no threshold seems to exist, health effects are likely to occur down to the background level of approximately $7 \mu\text{g}/\text{m}^3$ of PM10 (U.S. EPA, 1996).

3.4. Quality issues in air pollution data

The air quality monitoring networks produced by Project Automation are consistent with the current air quality regulations, at three level of network structure:

- Continuous analytical instrumentation
- Station acquirer
- Supervisor network centre

In particular, the following acts rule the monitoring of air pollution:

Instrumentation quality

Minister Act 2 April 2002, n. 60

The transposition of EU directive 1999/30/CE of 22 April 1999 on air quality limits the value for sulphur dioxide, nitrogen dioxide, nitrogen oxides, particles and lead and of directive 2000/69/CE on air quality limit value for benzene and carbon monoxide.

(Official Gazette n. 87, 13 April 2002- Ordinary Suppl. n.77)

Quality of monitoring networks**Government Act 4 August 1999, n. 351**

“Actuation of directive 96/62/CE on air quality evaluation and management matter”;

Minister Act 20 September 2002

Modes to guarantee the quality of the measurement system for air pollution, according to government act n. 351/1999.

(Official Gazette n. 231, 2-10-2002)

- Measures Quality guarantee → metrology
- Approval of sampling and measures equipment and definition of related procedures
- Accreditation of public and private laboratories for sampling and measures
- Approval of data collection networks

Technical documents:

- Criteria for EUROAIRNET of European Agency for the environment, February 1999, decision 2001/752/CE.
- Recommendations on the review of Council Directive 1999/30/EC-Draft 11-05-2004 of the CAFE Working Group on Implementation.
- Guidelines to the predisposition of air quality monitoring networks in Italy – APAT CTN ACE - 2004

Technical rules:

- **Rule EN 12341:1998** “Air quality - Determination of the PM10 fraction of suspended particulate matter - Reference method and field test procedure to demonstrate reference equivalence of measurement methods¹” approval by CEN ² in 2 November 1998 and transposed by UNI ³ in 6 April 2001.
- **Rule EN 14211:2005** “Ambient air quality – Standard method for the measurement of the concentration of nitrogen dioxide and nitrogen monoxide by chemiluminescence⁴” approval by CEN in 10 December 2004 and transposed by UNI in 2 September 2005.
- **Rule EN 14212:2005** “Ambient air quality – Standard method for the measurement of the concentration of sulphur dioxide by ultraviolet fluorescence⁵”, approval by CEN in data 10 December 2004 and transposed by UNI in 2 September 2005.
- **Rule EN 14625:2005** “Ambient air quality – Standard method for the measurement of the concentration of ozone by ultraviolet photometry⁶” approval by CEN in 10 December 2004 and transposed by UNI in 2 September 2005.
- **Rule EN 14626:2005** “Ambient air quality – Standard method for the measurement of the concentration of carbon monoxide by non dispersive infrared spectroscopy⁷” approval by CEN in 10 December 2004 and transposed by UNI in 2 September 2005.

¹ Air quality – definition of PM10 fraction of suspended particle material – Reference method and test procedure to demonstrate the equivalence of measures method with the reference method.

² European Committee for Standardization.

³ The Italian National Unification.

⁴ Air quality – Normalized method for the measure of concentration of nitrogen dioxide and nitrogen monoxide by chemiluminescence.

⁵ Air quality – Normalized method for the measure of concentration of sulphur dioxide by UV fluorescence.

⁶ Air quality – Normalized method for the measure of concentration of ozone by UV photometry.

- **Rule EN 14662-3:2005** “Ambient Air Quality – Standard Method for the measurement of benzene concentrations – Part 3: Automated pumping sampling with in situ gas chromatography⁸”, approval by CEN in 21 March 2005 and transposed by UNI in 6 October 2005.
- **Rule EN 14902:2005** “Ambient air quality – Standard method for the determination of Pb, Cd, As and Ni in the PM10 fraction of suspended particulate matter⁹” approval by CEN in 27 June 2005 and transposed by UNI in 20 October 2005.
- **Rule EN 14907:2005** “Ambient air quality – Standard gravimetric measurement method for the determination of the PM2.5 mass fraction of suspended particulate matter¹⁰” approval by CEN in 22 July 2005 and transposed by UNI in 29 November 2005.

Station Acquirer

Quality checks performed:

- On elementary data (sampled)
- On hourly data (elaborated)
- On equipment calibration

Validity of elementary data (samples, with a typical frequency of 5 sec.) is verified on the base of some parameters:

- Presence of anomaly in the equipment or station
- Acceptability lower/upper bound
- Difference between previous valid value of elementary data

Historical data (typically hourly mean values) are calculated from elementary valid data.

The validity of data is verified on the basis of some parameters:

- Number of elementary data used to calculate the excess of an upper threshold
- Difference between maximal value and minimal value of elementary data that contribute to the calculation of excess of a lower threshold
- Difference between maximal value and minimal value of elementary data that contribute to the calculation of excess of an upper threshold
- Historical data lower than a specific threshold
- Historical data upper than a specific threshold

Calibration of the analyzer:

- Zero verification: achievement of the regime of an equipment in presence of a null pollution concentration in atmosphere
- Span verification: achievement of the regime of an equipment in presence of a fixed pollution concentration in atmosphere

The validity of the values of zero and span calculated are checked on the basis of some criteria:

- Number of elementary data used for the calculation upper than a specific threshold
- Value calculated out of specific upper and lower acceptability thresholds
- Absence of equipment anomaly
- Difference with respect to the value calculated during the last valid calibration lower than a specific threshold

⁷ Air quality – Normalized method for the measure of concentration of carbon monoxide by IR spectroscopy not-dispersive.

⁸ Air quality – Normalized method for measurement of concentration of benzene – Part 3: Sampling for automatic pumping with gas-chromatography in situ.

⁹ Air quality – Normalized method for the definition of Pb, Cd, As e Ni in the PM10 fraction of suspended particle material.

¹⁰ Air quality – Normalized method for gravimetric measurement for the definition of the PM2, 5 mass fraction of suspended particle material.

In case of negative outcome of the first 3 validations, the calibration is considered not valid and warnings “zero defect” and “span defect” are produced.

In case of negative outcome of the last validation, the calibration is considered valid but diagnostic reports “zero check” and “span check” are produced.

Moreover are foreseen automatic diagnostic activities, which check the general operating status of the acquisition system, management of station equipment and storing of information on:

- operating status of I/O modules
- operating status of the system
- integrity status of archives
- correctness of configurations

Centre for network supervision

Data acquired by monitoring stations are manually validated by the operator. The application software offers tools for validation that allow operator to mark data as valid or invalid.

4. Quality of water data

Coastal waters are potentially affected by large and relatively constant inflows of contaminated water sources like streams, rivers or inadequately treated sewage, making monitoring essential to determine bathing water quality. Several physical, chemical and biological parameters are associated to water quality, from microbial quantification and water physical and chemical properties, to water quantity in streams, rivers and estuaries. In addition, weather conditions play an important role in bathing water quality, being relevant to bathing water fecal contamination. Several studies have shown that sunlight and air temperature have an important effect in microorganism die-off, therefore monitoring weather must also be considered.

In the Portuguese case study (Costa do Estoril) the streams that outflow to bathing waters are monitored by IST and SANEST, SA. Also, during summer season, bathing water quality is evaluated by the responsible national laboratory APA (Portuguese Environmental Agency) in accordance with the current legislation. To complement this evaluation IST/MARETEC has developed monitoring plans that include classic sampling in bathing waters and horizontal and vertical water quality profiles using sensors.

SANEST also developed a monitoring plan that includes the 14 streams of the Costa do Estoril, to evaluate streams water quality in different points of the watershed. This evaluation is important to determine streams possible impact in coastal waters. Complementary to this, a weather station (Guia WWTP), two hydrometric stations (streams), and one automatic acquisition buoy (Tagus upper estuary/river) acquire water quality data continuously. In addition, INAG, the Portuguese National Water Institute monitors water level and discharge in Portuguese rivers and water quality in surface and freshwaters, coastal waters and streams, publishing these and other useful results in a public internet site (<http://snirh.pt/>).

In the Dutch case study site the water data is gathered by several authorities. The majority of water quantity and quality data is collected via automated networks that belong to the waterboards in the area of the Province: Waterboard Brabantse Delta, Waterboard De Dommel and Waterboard Aa en Maas. These networks measure water levels, discharges, water quality parameters and ground water levels at various frequencies in time. The Province of North Brabant (PNB) has also measurements on water quantity, mainly on the deeper ground water levels.

4.1. Monitoring data

Water monitoring data can be acquired using 3 different sampling methodologies: Classic Sampling (discrete in time and space), Automatic stations (continuous in time and discrete in space) and Vertical and Horizontal Profiles (continuous in time and space);

4.1.1. Classic sampling

The sampling program required by the current legislation is a classic sampling program with fixed and predetermined sampling point in the bathing water. The classic sampling methodology is also used by IST and SANEST to evaluate bathing and streams water quality and quantity.

APA (Portuguese Environmental Agency) monitoring plan

This monitoring program is required by the current legislation: DL 236/98 (Portuguese transposition of 76/160/CEE Directive), and by the New Bathing Water Directive, that will repeal the current directive.

Sampling conditions: The water samples are collected in a fixed point (monitoring point) predefined in the monitoring plan. The sample is collected in appropriated bottles, supplied by the responsible laboratory. The water sample has to be kept cool and in the dark during transport to the analyzing laboratory.

Methods: The parameters are quantified in a certificated laboratory, following the standard methods defined by the legislation (Table 5).

Table 5. Parameters, Units and Methods required by the current legislation to Bathing Water Quality.

Parameter	Units	Method	Frequency/ Observations
Total Coliforms	CFU/100ml	Fermentation in multiple tubes. Subculturing of the positive tubes on a confirmation medium. Count according to MPN (most probable number) or membrane filtration and culture on an appropriate medium such as Tergitol lactose agar, endo agar, 0,4 % Teepol broth, subculturing and identification of the suspect colonies. In the case of 1 and 2, the incubation temperature is variable according to whether total or faecal coliforms are being investigated. Membrane-Filter Technique: MM 9.2; or Multiple-Tube Fermentation	Fortnightly (1)
Faecal Coliforms			
<i>Escherichia coli</i> *	CFU/100ml	Membrane-Filter Technique: ISO 9308-3 or ISO 9308-1	* Required by Directive 2006/7/CE
Enterococci (Faecal Streptococci) *	CFU/100ml	Membrane-Filter Technique: ISO 7899-1 or ISO 7899-2 Litsky method. Count according to MPN (most probable number) or filtration on membrane. Culture on an appropriate medium. Fermentation in multiple tubes. Subculturing of the positive tubes on a confirmation medium. Count according to MPN (most probable number) or membrane	* Required by Directive 2006/7/CE

		filtration and culture on an appropriate medium such as Tergitol lactose agar, endo agar, 0,4 % Teepol broth, subculturing and identification of the suspect colonies.	
<i>Salmonella</i>	/l	Concentration by membrane filtration. Inoculation on a standard medium. Enrichment — subculturing on isolating agar — identification.	(2)
Enteroviruses	PFU/10L	Concentrating by filtration, flocculation or centrifuging and confirmation	(2)
pH	-	Electrometry with calibration at pH 7 and 9	(2)
Colour	-	Visual inspection or photometry with standards on the Pt.Co scale.	Fortnightly (1) (2) No abnormal change in colour
Mineral Oils	mg/l	Visual and olfactory inspection or extraction using an adequate volume and weighing the dry residue.	Fortnightly (1) (2) No film visible on the surface of the water and no odour
Surface-active substances reacting with methylene blue	mg/l (lauryl-sulfate)	Visual inspection or absorption spectrophotometry with methylene blue.	Fortnightly (1) (2) No lasting foam
Phenols (phenol indices) mg/l C ₅ H ₅ OH		Verification of the absence of specific odour due to phenol or absorption spectrophotometry 4-aminoantipyrine (4 AAP) method.	Fortnightly (1) (2)
Transparency		Secchi's disc.	Fortnightly (1) (2)
Dissolved oxygen	% saturation O ₂	Winkler's method or electrometric method (oxygen meter).	(2)
Tarry residues and floating materials such as wood, plastic articles, bottles, containers of glass, plastic, rubber or any other substance. Waste or splinters.		Visual inspection.	Fortnightly (1)
Ammonia	mg/l NH ₄	Absorption spectrophotometry, Nessler's method, or indophenol blue method.	(3)
Nitrogen Kjeldahl	mg/l N	Kjeldahl method.	(3)
Other substances regarded as indications of pollution Pesticides (parathion, HCH, dieldrin)	mg/l	Extraction with appropriate solvents and chromatographic determination	(2)
Heavy metals such as: Arsenic Cadmium Chrome VI Lead mercury	As(2) Cd Cr VI Pb Hg	Atomic absorption possibly preceded by extraction	(2)
Cyanides	mg/l Cn	Absorption spectrophotometry using a	

		specific reagent	
Nitrates and Phosphates	mg/lNO3 PO4	Absorption spectrophotometry using a specific reagent	(3)

(1) When a sampling taken in previous years produced results which are appreciably better than those in this Annex and when no new factor likely to lower the quality of the water has appeared, the competent authorities may reduce the sampling frequency by factor of 2.

(2) Concentration to be checked by the competent authorities when an inspection in the bathing area shows that the substance may be present or that the quality of the water has deteriorated.

* Required by Directive 2006/7/CE

IST monitoring plan

In order to evaluate fecal contamination in streams outflow and bathing waters, IST developed a monitoring plan that considers sampling points near the coast line and in radial points. Streams discharge and water level is also measured. The surroundings of the Guia submarine outfall are also monitored, in 3 different water depths (surface, middle and bottom).

Measured parameters: measured parameters include Total Coliforms, Fecal Coliforms, *Escherichia coli*, Enterococci and Transparency.

Methods:

-Microbiological samples are collected to an appropriated bottle supplied by the laboratory, kept in cold and dark environment, and sent to the laboratory. To sample water at different depths a *Niskin bottle* is used. Total Coliforms, Faecal Coliforms, *Escherichia coli* and Streptococci (MPN/100ml) are quantified in laboratory following the standard methods mentioned in D.L. 236/98 and 2006/7/CE (Table 5).

-Transparency: water transparency is measured using a Secchi Disk. The disk is mounted on a pole or line, and lowered slowly down in the water. The depth at which the pattern on the disk is no longer visible is taken as a measure of the transparency of the water. This depth is known as the Secchi depth and is related to water turbidity.

- Level measurement (m): water level is measured in streams in a pre-defined location using a scale;

- Discharge measurement (m^3/s): Flow is measured near the stream outflow, in an area without sea influence, using a StreamPro ADCP or a Flow Tracker (both from Sontek).

SANEST monitoring plan

In order to determine streams impact on coastal water quality, SANEST assesses water quality in streams that outflow to Costa do Estoril bathing waters. A total of 30 sampling points distributed by the 14 existing streams are monitored.

Parameters studied: measured parameters include Total Coliforms (CFU/100ml), Fecal Coliforms (CFU/100ml), Intestinal Enterococci (CFU/100ml), Chemical oxygen demand (COD, mgO_2/l), Biochemical oxygen demand, 5-d, 20°C (BOD, 20°C, mgO_2/l), Ammonia ($mgNH_4^+/l$), Dissolved Oxygen (DO, $mg.l^{-1}$), pH, Temperature (°C) and Conductivity ($\mu S cm^{-1}$).

Methods: sample collection, conditioning and transport are performed following method 1060 of the “Standards Methods for the Examinations of Water and Wastewater”. Analysis methods follow the standards required by the Portuguese DL 236/98, presented in Table 5.

4.1.2. Automatic stations

In the scope of the monitoring programs, different types of data are continuously collected by automatic systems. Monitoring plans include two hydrometrics stations in streams, one automatic acquisition buoy in the upper Tagus estuary and one meteorological station in Guia WWTP.

Meteorological Station: Property of SANEST, located in Cascais, on the top of Costa do Estoril Wastewater Treatment Plant (38°41'41''N, 9°26'48''W). This station measures temperature (°C), humidity (%), wind direction (°) and intensity (m/s), atmospheric pressure (mBar), solar radiation (W/m²), and precipitation (mm/30min). All the sensors are connected to a Campbell Scientific CR10X Datalogger. All the recorded data is remotely sent to a central workstation in IST using a GSM Siemens TC35T modem.

This system is supported by solar energy that powers batteries. If the battery decreases above a fixed value an alarm is set. Instruments include a temperature and humidity sensor 50Y from CS; Atmospheric Pressure is measured with a *Druck* RPT410F, solar radiation with a *Kipp & Zonen* CM3, precipitation by an ARG100 udometer and wind velocity and direction are measured by an A100R anemometer and a Wind Van W200P from *Vector Instruments*, respectively.

Hydrometric Stations- In Costa do Estoril two automatic stations were installed, in two streams, near its outflow but without sea influence. These stations measure the water level with a pressure sensor, and water temperature using an integrated NTC-temperature sensor. The data is acquired continuously, recorded systematically, and stored in a Gealog Compact Datalogger. Recorded data is sent by a Gealog GSM modem EGSM900/1800, to the computer workstation in IST. These stations have the capability to send alarms when the water level rises (or decreases) a determined threshold value. The equipment is powered by a 12 V battery.

Automatic Acquisition Buoy An autonomous recording system for continuous, long-term, water quality and current in-situ monitoring, has been installed in 2007, at the upper Tagus Estuary. The so-called SIMPATICO system is composed by various elements that include a surface floating buoy, a datalogger, a multi-parameter sonde for water quality monitoring, and an acoustic doppler current meter profiler (ADP), deployed at distance from the buoy.

The multi-parameter sonde (YSI 6600 V2-4) is equipped with optical sensors to measure turbidity (NTU), saturated (%) and dissolved oxygen (DO, mg/l) and chlorophyll (µg/l, via fluorescence), plus (non-optical) sensors for conductivity (µS/cm), temperature (°C) and pH. Salinity is determined automatically from the conductivity and temperature readings.

The acoustic current profiler (SONTEK Argonaut SL) is fitted with a built-up pressure sensor, for water level variations, and velocity flow measurements (plus compass/tilt and temperature sensors). The Argonaut-SL (side-looking) measures 2D currents in an adjustable measurement volume located at a range up to 120 m. It uses two acoustic transducers with slant angle of 25 degrees to record the flow velocity in a plane parallel to the water surface, and an additional vertical beam for stage (i.e. water level) measurement.

The recorded data are remotely downloaded via cellular telemetry at scheduled intervals, and directly integrated into a database with web access. This system provides daily online water quality (temperature, salinity, pH, dissolved oxygen, turbidity and chlorophyll) and currents and flow data and is available at <http://www.mohid.com/tejo-op/>.

4.1.3. Profiles using sensors

Estuaries are highly variable in space. Horizontal and vertical variability of physical and chemical parameters can be monitored by performing profiles using sensors. This method allows studying several parameters in the entire water column identifying water homogeneities and heterogeneities.

Multi-parameter sonde (YSI 6600 V2-4): The 6-series environmental monitoring systems are multi-parameter water quality measurement and data collection systems intended for use in research, assessment and regulatory compliance applications. This probe is equipped with optical sensors to measure turbidity (NTU) and chlorophyll ($\mu\text{g/l}$, via fluorescence), plus (non-optical) sensors for conductivity ($\mu\text{S/cm}$), temperature ($^{\circ}\text{C}$), saturated (%) and dissolved oxygen (DO, mg/l) and pH. Salinity is determined automatically from the conductivity and temperature readings.

Acoustic Doppler Current Profiler (WorkHorse Sentinel ADCP from SONTEK) - Using Doppler Effect this device can measure water velocity (m/s) and direction ($^{\circ}$) in the entire water column, or in defined layers.

Current Meter (Aquadopp from Nortek) – this equipment measures current velocity (m/s) using Doppler technology. This device measures only one water layer from 1 to 4 meters high.

Mapping System In order to capture horizontal variability, an automatic acquisition system combining sensors with GPS positioning has been developed by the Technical University of Lisbon (IST/Maretec). The so-called “Mapping System” was designed to obtain continuous water quality measurements from a moving vessel. Basically, in this system, real-time probe measurements are integrated (using a datalogger) with data from a flow tracker and a GPS, thus obtaining a horizontal profile of the studied area.

The multi-parameter sonde used is the YSI 6600 V2-4. Other components of the system include a system of tubes and a water pump, powered by a 12V battery. In the data circuit, the datalogger integrates the data from the sonde, GPS and flow tracker, which is later downloaded to a PC or pocket PC. The system allows real-time visualisation of results, using windows based software provided by the manufacturer. After laboratory tests, this equipment has been used successfully in different study areas such as Tagus, Guadiana, Arade and Almarem estuaries, Óbidos coastal Lagoon and Ria Formosa.

4.2. Reference values in water quality

The current legislation applied to bathing water quality is the 76/160/CEE Directive, transposed to the Portuguese legislation as DL 236/98 from 1 of August. Directive 2006/7/EC- New Bathing Water Directive will repeal Directive 76/160/EEC with effect from eight years after the entry into force.

Council Directive of 8 December 1975 concerning the quality of bathing water (76/160/CEE), transposed to Portugal as D.L.236/98: This directive considers 19 parameters for analysis. Some of these have to be quantified fortnightly, but others just have to be checked if an inspection in the bathing area shows that the substance may be present or that the quality of the water has deteriorated. This directive proposes a guide and a mandatory value for the bathing water fecal contamination (Table 6).

Directive 2006/7/CE of the European Parliament and of the Council of 15 of February 2006- The New Bathing Water Directive considers 3 fecal parameters, but requires an evaluation of all the bathing water pollution sources, through the Bathing Water Profile. The bathing water can be classified as “poor”, “sufficient”, “good”, or “excellent”. The bathing water evaluation considers the 3 previous bathing seasons and is based upon a 95 or 90 percentile evaluation (Table 7).

For streams water quality there is not a European legislation, however, INAG, Portuguese Water Institute, defines **Water Quality Limits for Superficial Waters of Multiple Uses** (Table 8).

Table 6. Directive 76/160/CE quality requirements to bathing water quality from 76/160/CE.

Parameter	Units	Guide Value	Imposed Value
Total Coliforms	CFU/100ml	500	10000
Faecal Coliforms	CFU/100ml	100	2000
Enterococci (Faecal Streptococci)	CFU/100ml	100	
<i>Entero viruses</i>	PFU/10L	-	0
<i>Salmonella</i>	/l	-	0

Table 7. Directive 2006/7/CE quality requirements to bathing water quality

Parameter	Units	Excellent Quality	Good Quality	Sufficient Quality
Intestinal Enterococci	CFU/100ml	100 (*)	200 (*)	185 (**)
<i>Escherichia coli</i>	CFU/100ml	250 (*)	500 (*)	500 (**)

(*)- Based upon a 95-percentile evaluation.

(**) Based upon a 90-percentile evaluation.

Table 8. Portuguese classification of water quality in superficial waters of multiple uses

Parameter	Class	A	B	C	D	E
		Without Pollution	Weakly Polluted	Polluted	Very Polluted	Extremely Polluted
pH		6.5-8.5	-	6.0-9.0	5.5-9.5	5.0-10.0
Conductivity	$\mu\text{S}/\text{cm}$ 20°C	≤ 750	751-1000	1001-1500	1501-3000	> 3000
Total Suspended Solids (TSS)		≤ 25	25.1-30.0	30.1-40.0	40.1-80.0	> 80
Dissolved Oxygen (DO)	%	≥ 90	89-70	36-50	49-30	< 30
Biochemical oxygen demand, 5-d, 20°C (BOD, 20°C)	mg O ₂ /l	≤ 3	3.1-5.0	5.1-8.0	8.1-20.0	> 20.0
Chemical oxygen demand (COD)	mg O ₂ /l	≤ 10.0	10.1-20.0	20.1-40.0	40.1-80.0	> 80.0
Oxidability	mg O ₂ /l	≤ 3	3.1-5	5.1-10	10.1-25	> 25
Free ammonia	mg NH ₄ /l					
Nitrate	mgNO ₃ /l	≤ 5.0	5.0-25.0	25.1-50.0	50.1-80.0	> 80.0
Nitrite	mg NO ₂ /l	≤ 0.01	0.011-0.020	0.021-0.15	0.16-0.3	> 0.3
Total Coliforms		≤ 50	51-5000	5001-50000	> 50000	-
Faecal Coliforms		≤ 20	21-2000	2001-20000	> 20000	-
Faecal Streptococci		≤ 20	21-2000	2001-20000	> 20000	-

Iron (Fe)	mg/l	<=0.50	0.51-1.00	1.10-1.50	1.50-2.00	>2.00
Manganese (MN)	mg/l	<=0.10	0.11-0.25	0.26-0.50	0.51-1.00	>1.00
Zinc (Zn)	mg/l	<=0.30	0.31-0.05	1.01-3.00	3.01-5.00	>5.00
Copper (Cu)	mg/l	<=0.020	0.021-0.05	0.051-0.200	0.0201-1.000	>1.00
Chromium (Cr)	mg/l	<=0.010	-	0.011-0.050	-	>0.050
Selenium (Se)	mg/l	<=0.005	-	0.0051-0.010	-	>0.010
Cadmium (Cd)	mg/l	<=1.0	-	1.1-5.0	-	>5.0
Lead (Pb)	mg/l	<=0.050	-	0.051-0.100	-	>0.100
Mercury (Hg)	mg/l	<=0.5	-	0.51-1	-	>.
Arsenic	mg/l	<=0.010	0.011-0.050	-	0.051-0.100	>0.100
Cyanic	mg/l	<=0.010	-	0.011-0.050	-	>0.050
Phenol	mg/l	<=1.0	1.1-5.0	51-10	11-100	>100
Surfactants	mg/l	<=0.2	-	0.21-0.50	-	>0.5

5. Data quality management

As introduced in the previous paragraphs, the collection of environmental data is performed in several scenarios, 2 for air data and 3 for water data:

- i. **Air pollution monitoring networks:** a relatively small number of “large stations”;
- ii. **Sensor networks:** typically Wireless Sensor Networks (WSN) ;
- iii. **Classic sampling** - discrete variability in time and in space;
- iv. **Automatic stations** - continuous variability in time, discrete in space;
- v. **Profiles/ Mapping System-** continuous variability in time and space;

In the following paragraphs the different data quality issues that arise from each of these scenarios will be analyzed.

5.1. Data quality management in air pollution monitoring networks

In the case of centralized data, the poor quality of sensor data due to limited sensor precision as well as sensor failures and malfunctions has to be managed, since decisions derived on incorrect or misleading sensor data are likely to be faulty. The issue of how to efficiently provide applications with information about Data Quality (DQ) is still an open research problem.

The solution proposed in [4] consists in maintaining a flexible number of DQ dimensions together with the data, and propagate them in support of applications directly consuming streaming data or processing data filed in a persistent database. The propagation of data quality information results in an overhead for data transfer and management. Due to the large amount of stream data, this may shape up as very expensive. Furthermore, quality information presents additional metadata on sensor data. Yet, the management of data quality is addressed neither in data stream nor in relational metadata models. In sensor networks, where there is a distributed processing capability,

i.e. some computations can take place at the device level, offer a solution by which quality and volatility of the data can be managed at the device level, while queries are sent to a transactional data warehouse where you can perform probabilistically accurate queries.

A sensor data stream D consists of n attributes A_i ($1 \leq i \leq n$) representing sensor measurements. In the traditional metadata model, each attribute A_i is associated with an unrestricted number of data value items v_{ij} . The output of a sensor is a discretized and digitized data stream representing the measured physical values. The characteristics of the sensor define the data quality dimensions of the outgoing data stream. Without loss of generality, we focus on the two important DQ dimensions accuracy and completeness, which are calculated on the sensor data stream. The measured sensor data is streamed towards the target applications, where the data is processed and actions or decisions are derived. The naive approach of data quality annotations consists in streaming the data quality information for each DQ dimension (grey) with the same stream rate as the measurement stream (white) as shown in the next figure. The data item is not only defined by its numerical values, but further described by its DQ information Fig. 3. Hence, this approach is not suitable for those applications with stringent resource constraints and should only be employed in case that communication costs for data transmission are not significant.

To reduce the additional data volume to transfer data quality information in a data stream, in [4] is proposed the usage of jumping data quality windows: the data quality information is not sent together with every single data item but window-wise for each DQ dimension. The notion of jumping windows is interposed in the relation between attribute and data item as shown in the next figure (Fig. 4).

Each measurement attribute stream is parted in an unlimited number of windows of a given size s containing sensor data items (white) and data quality information (gray). Each window is identified by its starting point $t_{begin} = t_k$. It consists of s measurement values v_{ij} ($k \leq j \leq k + s - 1$) of a certain attribute A_i . Furthermore, the window contains one value for each DQ dimension q_{ik} (e.g. window completeness c_{ik} and window accuracy a_{ik}). The additional memory space S to cover d_i data quality dimensions for each of n attributes A_i depends on the attributes' window size s_i and the stream length m .

$$S = m \cdot \sum_{i=1}^n \frac{d_i}{s_i}$$

The insertion of data quality information in relational databases is obtained at the metadata level: every column in a relational table is enhanced with d data quality characteristics (a.k.a. DQ dimensions). Important thereby is the maintenance of the resource-saving window model, so that data quality information is not stored for every measurement value v_{ij} . Therefore, the database table containing sensor data is partitioned into relation windows analogue to the jumping stream windows. A **Table** or **View** is composed as a **ColumnSet** of a given number of **Columns**, describing the table (or view) attributes. A **Row** represents an instance of a certain **ColumnSet** including the inserted data values. The relational window to manage data quality information can be configured as a **RowSet** containing the sensor data of a certain time interval. The **Data Quality** associated to a certain **Column** is stored in reference to specific **RowSets** of the corresponding **ColumnSet**. As exemplary data quality dimensions Fig. 5 shows the **Accuracy** and **Completeness**. In the context of centralized data, software tools have also been proposed to improve the quality of the data after they are stored into a database. Table 9 presents some of them.

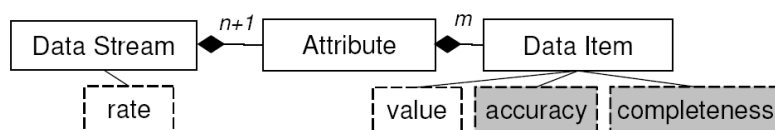


Fig. 3. Inclusion of data quality measures in the data.

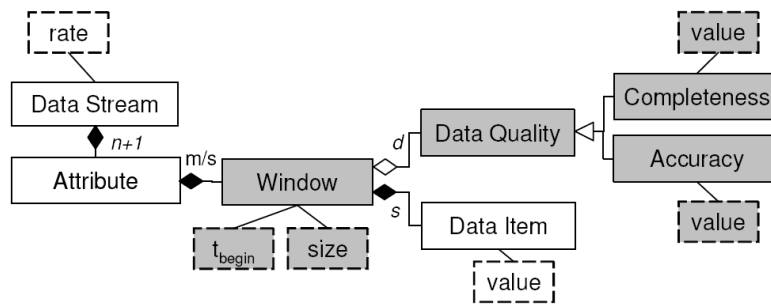


Fig. 4. Data quality information estimated on a jumping data window.

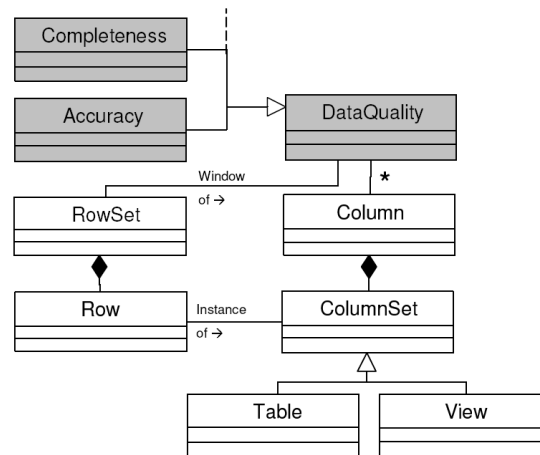


Fig. 5. Measures of data quality: Completeness and Accuracy.

Table 9. Software tools proposed to improve the quality of the data.

Name	Activities	Features	Application domain
Potter's wheel	Standardization. Object identification and deduplication. Data integration – instance-level conflict resolution. Profiling (structure extraction)	Tightly integrates transformations and discrepancy/anomaly detection	[General purpose]
Telcordia's tool	Standardization. Object identification and deduplication	Record linkage tool parametric wrt distance functions and matching functions	Addresses. Tax-payers and their identifiers
Ajax	Object identification and deduplication. Data integration – instance-level conflict resolution.	Declarative language based on logical transformation operators. Separates a logical plan and a physical plan	Bibliographic references
Artkos	Standardization. Data integration – instance-level conflict resolution. Error localization	Covers all aspects of ETL processes (architecture, activities, quality management) with a unique metamodel	ETL process of an enterprise DW for health applications and pension data
Choice Maker	Object identification and deduplication	Uses clues that allow rich expression of the	People's/business names and addresses. Medical

		semantics of data	data. Financial/credit card records
Intelliclean	Object identification. Choice of representative object	Uses two types of rules, for object identification and object merging	[General purpose]

5.2. Data quality management in wireless sensor networks

The quality management in wireless sensor networks is addressed in [1] and [2]. These works fit into the research field of in-network data fusion, and partly follows a strategy more recent than sensor data clustering, based on the use of replicated and distributed *data models* among the network nodes, cluster leaders and base stations.

The idea underlying this approach is to transmit over the network just notifications and information about notable events and keep the replicated models in an aggregation node (sink), to avoid most of the time the transmission of raw samples thus lowering the use of the radio connection and the energy consumption as well. In these scenarios, simple rather than heavy and centralized models can be executed also in devices with limited capabilities and let build applications that are more flexible with respect to variation in the data distribution.

The paradigm of “in-network aggregation” essentially consists in the exploiting of the (limited) computational power of the network nodes to implement data fusion algorithms *while* the data travels around pointing to the base station. In fact, when dealing with data distribution, a first common approach for learning a model is to gather all the data to a central site and then execute any algorithm developed for the centralized case. A more efficient approach, which consists in splitting the procedure in a *local learning* and *global learning* phase, i.e. executing some processing at each remote site (close to sensors) and combining the results without transmitting raw data.

The general organization of such system comprises 3 parts:

1. A set of *sensor nodes*, which continuously monitor a set of attributes $F = (F_1, F_2, \dots, F_m)$ each of which evolves over time. Each node computes its own local model for prediction and filtering purposes. This model exploits only temporal correlations. As new readings arrive, the running model is checked for “plausibility”, in statistical terms; if it fails the test, a new model is computed and the new set of parameters is notified to the base station.
2. A *base station (sink)*, which maintains the set of parameters of the local model estimated by each *sensor node* and can be queried, in a probabilistically controlled way, about values at each node, without requiring communication over the network.
3. A *global model*, also maintained at the base station, which can learn and exploit the spatial and the inter-attribute relationships and answer queries about the “state of the world” performing network-wide inference.

In this local/global model scenario, the check of data quality is performed online at the sensor node, simulating the evolution of the data generated by the local model and comparing them with real data. This allows sending (push) information to the sink only when an anomalous data is detected. The detection is targeted on two types of anomalies: *outliers*, which occur when the sensor reads a value outside a given confidence interval, and *changes in the distribution* of the observed data, when repeated errors in forecast require a model update. In the first case is sent a notification

message for each outlier, the second case is limited to the transmission of the parameters of the updated local models.

Model update has computation and communication requirements, and hence a substantial power consumption; it has to be executed only as a consequence of repeated errors in the forecast. For this purpose it is applied a monitoring process based on multiple error thresholds. It is proven in [5] that given a constant $v \geq 1$ and calculating a threshold ε , the prediction $P(t)$ for time t is in the interval $[P(t)-\varepsilon \leq v_i \leq P(t)+\varepsilon]$, centred on the “true” value v_i , with error probability at most $1/v^2$. To increase the sensitivity on errors, and better control the update of the model, it is chosen a second threshold $\delta < \varepsilon$: a model is a good predictor if the absolute error in forecast falls in $[0, \delta]$, if the error is repeatedly between δ and ε the model might be updated; if error exceeds ε the new sample can be considered an outlier (produced by e.g. a damaged sensor).

The monitoring process starts after model estimation: at each time step t the sensor forecasts a value and compares it with the new data collected. A queue of Λ subsequent readings is maintained, called *monitoring window*: the model is updated if in this window the absolute error between real value and forecast exceeds δ (including outliers) for more than a percentage a of the monitoring window size.

This process of statistical analysis can be applied to produce accuracy indicators e.g. rate of data exchange or reliability of the device.

5.3. Data quality management in water data

As described in paragraph **Errore. L'origine riferimento non è stata trovata.**, different ways of water quality data collection are being used that originate a large multiparameter dataset which can be discrete in time and space, discrete in space and continuous in time, or continuous in time and space. These different results require different data management proceedings. IST has developed different data management methodologies, according to each type of data.

5.3.1. Data quality management to classic sampling results

IST developed a database to store classic sampling measurements, where available data can be found in <http://mohid.com/gis/>. All data is centralized in a POSTGRESQL database. To upload data to the database, all data pass by several different steps, some of them helping to identify gaps and errors. The sampling results are placed in Excel Bulletins, and uploaded to the Database, using LabManager (a stand-alone backoffice application to upload data to database, developed by IST).

Before being sent to LabManager, the excel bulletins are reviewed by an application that confirms that all the data fields are with values or names. This application also identifies incorrect data format (i.e. confirming that all samples have an associated laboratory, result and observation field fulfilled). In these bulletins, it is necessary to fill different fields: sample point; sample point reference; sampling date; sampling time; sample ID; depth; observations; parameters results; laboratory responsible for the analysis.

The bulletin is then loaded in LabManager application, where information about the monitoring project responsible for that sampling is added. This is an important field to understand results, because each monitoring plan has its objective and study case. Lab Manager Application uploads the new values into the GestãoCosteira DataBase (Coastal Management DataBase), and the values are automatically online.

5.3.2. Data quality management in automatic stations data

Data from automatic stations is sent to the central workstation in IST through GSM communication. Then, an application *ReadText*, developed by IST, reads ASCII data and writes it into a Microsoft Access. An alert system to control automatic stations data quality, integrity and communication procedures was also developed by IST- *StationWatcher*. This application monitors if the data arrival to the database is in a pre-defined time period; checks all data parameter values inside a pre-defined range; and sends emails to pre-defined users alerting about unexpected data values or missing data. Examples are presented in Table 10 and Table 11.

Table 10. StationWatcher filter to Guia Meteorological Station data:

Parameter		Range	
		Minimum	Maximum
Hours delay		--	32
Solar Radiation	W/m ²	0	1050
Wind velocity	m/s	0	30
Wind Direction	°	0	360
DP Wind Direction	°	0.002	75
Atmospheric Pressure	mBar	950	1050
Temperature	°C	-5	48
Humidity	%	5	100
Datalogger Temperature	°C	-5	50
Environmental box	°C	-5	50
Battery	V	10	15

Table 11. StationWatcher filter to Tagus Buoy data

Parameter		Range	
		Minimum	Maximum
Hours delay		--	32
Water Temperature	°C	6	30
Conductivity	µS/s	0	1
Salinity	PSU	0.01	0.5
pH		6.5	9
Turbidity	NTU	1	200
Chlorophyll	µg/l	0	200
Saturated oxygen	%	70	150
Dissolved oxygen	mg/l	7	12
Battery	V	10	15

5.3.3. Data quality management in profiles

Profile Manager is the data management software created by IST to load data from horizontal and depth profiles to a database. This software can load data from: ADCP in a fixed point; ADCP transect; YSI sonde (without datalogger); YSI sonde (with datalogger/mapping system); Mohid GIS format (.srm timeserie); to a Database.

This software has three main application areas:

- 1- Graphical Interface
- 2- Data Base
- 3- Folders management system

1- The graphical interface allows the interaction between the user and the data collected. The user can: save the data collected, or upload it directly from the acquisition device; view and change data; filter and select the data before upload them into the database (Fig. 6). This application is able to produce graphic results using Mohid GIS (IST application), but this feature is still under development. The aim is to produce graphical information in graphs, or directly in web services. The development platform chosen was .NET2005, using VB.NET computer language (**Errore. L'origine riferimento non è stata trovata.**).

2- Data Base

The Database is prepared to receive several data. Before upload, data must be filtered, to avoid load of invalid data into the database. **Errore. L'origine riferimento non è stata trovata.** shows a view of several data relations in the database.

3- Folders Management

Once database contains only valid data, it is useful to save the original folders in other partition. This is also done by ProfileManager.

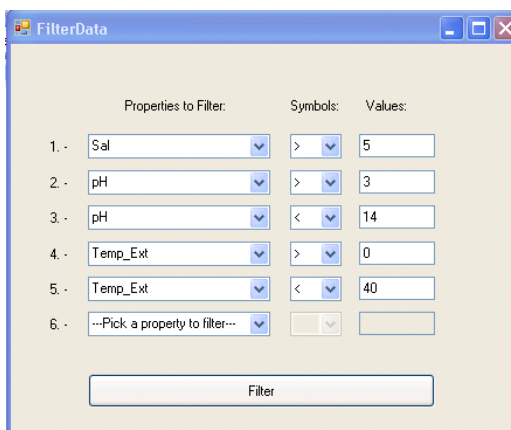


Fig. 6. FilterData

The screenshot shows the Profile Manager application window. At the top, there are menu options (File, Tools) and buttons for 'Load From File', 'Load From DB', 'Load From Timeserie', 'Filter Data', 'Save As MOHID GIS', and 'Add to DB'. Below these are 'Load From File', 'Load From DB', and 'Load From Mohid TimeSerie' buttons. The 'Profile Details' section contains fields for Name, Type (Sonda YSI), Equipment, Plane, Minimum/Maximum coordinates (XX, YY, ZZ), Project, Samples (1352), Begin/End times, Operator, and File path. A 'Tips' box on the right suggests adding data to a database. The main area is a table with columns: Instant, XX, YY, ZZ, CorGPS, NSat, Temp1, Conduct, Sal, O2, O2_tot, pH, ORP, Turb, Clor, Bat, Temp2. The table contains 20 rows of data for the date 2007-04-11.

Instant	XX	YY	ZZ	CorGPS	NSat	Temp1	Conduct	Sal	O2	O2_tot	pH	ORP	Turb	Clor	Bat	Temp2
2007-04-11 10:15:05	-9.45868	38.69249	0.034	1	7	14.81	47.68	31.07	93.6	7.83	8.02	151.3	31.5	4.1	12.3	14.81
2007-04-11 10:15:15	-9.45907	38.69246	0.032	1	6	14.8	49.02	32.04	93.4	7.77	8.02	151.5	31.5	4.1	12.3	14.8
2007-04-11 10:15:25	-9.45947	38.69246	0.034	1	5	14.81	46.65	30.32	93.4	7.85	8.02	151.3	31.5	4.1	12.3	14.81
2007-04-11 10:15:35	-9.45988	38.69245	0.028	1	6	14.79	45.49	29.49	93.2	7.88	8.01	151.8	24	3.8	12.3	14.79
2007-04-11 10:15:45	-9.46031	38.69246	0.027	1	6	14.78	47.19	30.71	93	7.8	8.02	151.8	15.1	3.3	12.3	14.78
2007-04-11 10:15:55	-9.46074	38.69255	0.025	1	7	14.78	47.22	30.73	92.8	7.78	8.01	152.2	17.1	4.3	12.3	14.78
2007-04-11 10:16:05	-9.46116	38.69259	0.025	1	6	14.79	48.07	31.35	92.7	7.74	8.01	152.1	17.2	3.9	12.3	14.79
2007-04-11 10:16:15	-9.46158	38.69256	0.024	1	7	14.82	48.13	31.4	92.5	7.73	8.01	152.4	27.1	4.4	12.3	14.82
2007-04-11 10:16:25	-9.46196	38.69249	0.025	1	7	14.82	41.52	26.65	92.4	7.94	8.01	152.4	25.4	4.4	12.3	14.81
2007-04-11 10:16:35	-9.4624	38.69251	0.023	1	7	14.82	48.01	31.31	92.3	7.71	8.01	152.7	21.9	4	12.3	14.82
2007-04-11 10:16:45	-9.46282	38.69262	0.024	1	7	14.83	48.81	31.89	92.2	7.67	8.01	152.7	18.9	4.4	12.3	14.84
2007-04-11 10:16:55	-9.46323	38.69267	0.024	1	7	14.85	47.64	31.04	92.2	7.71	8	153.1	23.9	4.1	12.3	14.85
2007-04-11 10:17:05	-9.46366	38.69271	0.025	1	6	14.85	43.22	27.86	92.1	7.86	8	152.9	25.6	3.9	12.3	14.85
2007-04-11 10:17:15	-9.4641	38.69271	0.024	1	7	14.88	47.3	30.8	92.1	7.71	8	153.3	33.7	3.6	12.3	14.88
2007-04-11 10:17:25	-9.46451	38.69272	0.025	1	5	14.89	47.13	30.68	92.1	7.71	8	153.3	30.3	4.1	12.3	14.89
2007-04-11 10:17:35	-9.46494	38.69276	0.024	1	6	14.85	47.59	31.01	92	7.7	8	153.7	23.5	3.9	12.3	14.85
2007-04-11 10:17:45	-9.46534	38.69281	0.025	1	5	14.84	47.97	31.28	92	7.68	8	153.7	22.8	3.8	12.3	14.84
2007-04-11 10:17:55	-9.46577	38.69284	0.023	1	5	14.87	48.08	31.36	92	7.68	8	154	22.2	2.9	12.3	14.87
2007-04-11 10:18:05	-9.46621	38.69286	0.024	1	7	14.89	43.26	27.89	92	7.83	8	154	23.8	2.9	12.3	14.89

Fig. 7. Inside view of Profile Manager

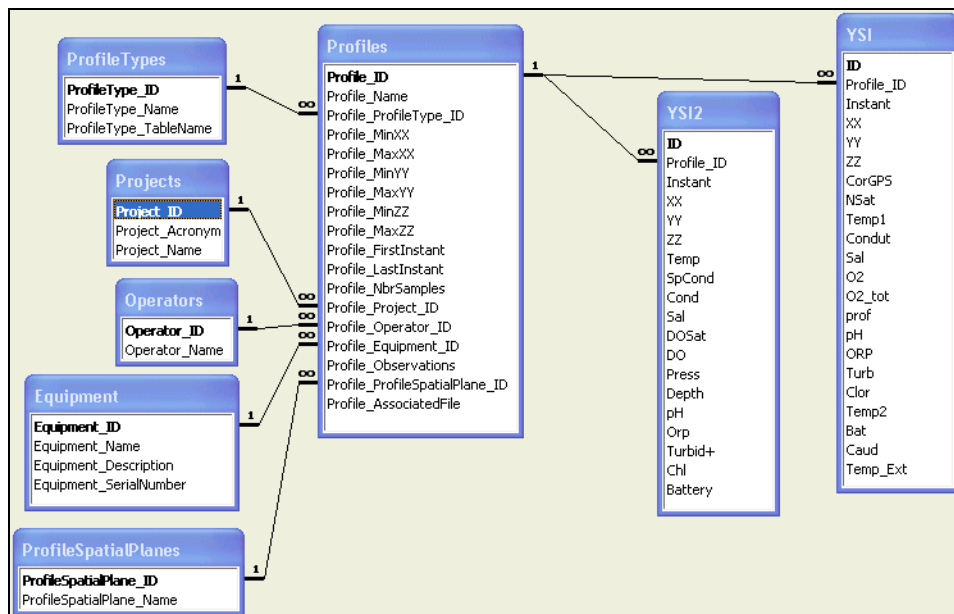


Fig. 8. Example of relation between fields in the database.

5.3.4. MetaDatabase

The metadata kept in the database are saved in different fields: Name, type, equipment, geographic location, project, instant period, operator, number of samples, observations, original data file path. Each dataset must answer to some questions (Table 12):

- Where? Where was collected/produced the data.

- When? When was the data collected/produced.
- What? Which parameters were collected or simulated?
- How? How was the data collected or simulated?
- Who? Who is responsible for this information?
- Where can be found? Where can the information be found?
- Other information: Data format, description, information resume, etc.

Table 12-Data sets with information about sample collection to answer a several questions: Where; When; What; How; Where can be found.

Where?

• Field	• Comment
<i>North</i>	North Coordinate (Mercator Projection)
<i>South</i>	South Coordinate (Mercator Projection) (must be equal to the North coordinate if it's a discrete sample).
<i>East</i>	East Coordinate (Mercator Projection) (must be equal to the North coordinate if it's a discrete sample)
<i>West</i>	West Coordinate (Mercator Projection)
<i>Measuring area type</i>	Point (Punctual), Track or Area observation.
<i>Coordinate system</i>	Define the coordinate's type. Normally are use geographic coordinates with Datum = WSG 84.
<i>Water depth</i>	Depth of sample collection, Can be: Surface, Bottom, Multi Level or Not Applicable.
<i>Vertical datum</i>	Reference lower water standard

When?

• Field	• Comment
<i>Start date</i>	Beginning of the observation.
<i>End date</i>	End of the observation.
<i>Sampling Interval</i>	Time between consecutive samples.

What?

• Field	• Comment
<i>Parameter</i>	Was created a hierarchy by observation area to better identify the parameter type (e.g. Water Body, Land, Atmosphere, Socioeconomics). This field helps to a better user search, better search all the data
<i>Abstract</i>	Short description about observation

How?

• Field	• Comment
<i>Method</i>	Method in use
<i>Instrument</i>	Instrument model
<i>Platform</i>	Type of platform where was realized the sample

Where can be found?

• Field	• Comment
<i>Originator</i>	Organization responsible for the sample
<i>Data reference</i>	Specific Reference (ID) (if exist)
<i>Database reference</i>	Identification of the database where the data is stored.
<i>Name by which the data set is known</i>	Name of the monitoring program.
<i>Short name of data set</i>	Monitor Program acronym.
<i>Access/ordering of data</i>	Type of access to data.

<i>Internet access/ordering</i>	URL to data access.
<i>E-mail data contact</i>	Email of the data provider or responsible
<i>Restrictions on access to the data</i>	Restrictions to data access.

5.4. Water quantity data management

Water quantity data collected by automatic networks can be processed using classification with data labelling. At several Dutch water boards WIS (Waterkwantiteits Informatie Systemen, Waterquantity Information Systems) are operated which store, process, validate and label the water quantity data. In these systems raw data is imported in a data storage and stored for archival purposes. After the raw data storage a classification procedure will assign data quality labels for each individual value in the timeseries data.

In this classification process we distinguish *primary validation* and *secondary validation*. *Primary validation* handles data quality errors due to for example measurement errors in the sensors being used or errors in data communication from the sensor to the data storage. The criteria for quality labelling are a given for each variable at each location and typically incorporate: a minimum allowed value, maximum allowed value, maximal value increase per timestep and maximal value decrease per timestep or missing value. The raw data timeseries are labelled using these criteria and the labels are stored with the time series data.

Table 13. Water quantity data qualification labels used in WIS-Waterkwantiteits Informatie Systemen (Waterquantity Information Systems) at several waterboards in the Netherlands.

Label	Description
P	Primary valid
L	Below minimal allowed value
H	Above maximal allowed value
Ds	Maximal increase per timestep exceeded
Dd	Maximal decrease per timestep exceeded
M	Missing
S	Static: no change in at least n timesteps

In case a value is not primary valid ('P') the raw value can be modified using:

- Automated interpolation of missing values (label 'M') using linear or block interpolation for a maximum of timesteps t-n
- Automated change of extreme values using (label L of H) using linear or block interpolation for a maximum of timesteps t-n

The number of timesteps n for automated interpolation of values is set per timeserie. Interpolated values are stored in new records together with their labels ('I' for Interpolated), the raw data is not overwritten.

After the *primary validation*, *secondary validation* can take place. This validation is done on both primary valid as primary invalid data, and can be one of the following processes.

1. Both primary valid and invalid data can be corrected using linear regression modelling, which will give an estimation of values in a timeserie. This type of secondary validation is also known as reconstruction of data, and is can be done to fill in missing values using different time steps from the time serie or using other locations in the vicinity. This method can be performed in case of short periods with missing or invalid data.

2. In case there is a systematic error in a timeserie, for example due to an incorrect calibrated sensor, the timeserie can be corrected by means on increasing or decreasing the values in the timeserie using a constant or even variable factor in time.
3. Finally time series can be validated and corrected using water balances. With a water balance formula several locations can be incorporated in the calculation and the values at the locations can be checked for correctness.

Secondary validation can be performed automatically. Each newly calculated data value has a label which corresponds to the formula with which it is calculated.

6. Data quality requirements from users

The different lenvis user groups: i) Professional users; ii) Public users; and iii) Providers, defined in D 1.2, have different objectives for using the data and also different ways to provide it. However, data used in the lenvis system will have to be acquired, managed and loaded considering the proceedings explained in section 2, 3, 4 e 5, which guarantee that the data is validated.

Professional users, as well as government bodies, organizations and public institutions must follow procedures required by European Community and international regulations regarding environmental data quality, as, for example, protocols like the European Information and Observation Network (Eionet), briefly described in D 1.3. lenvis will also follow these requirements, so that data quality requirements from the user side are in line with the ones in the project.

However, other regulations, at national or regional levels, may exist and may being used by potential lenvis users that have not been identified at this time. Also, new requirements arising from either new international or national regulations or even internal standards from the inside of institutions may come up. Therefore, lenvis will have to pay particular attention to this fact during the development of the project and verify if no other requirements from the users' side regarding data quality will emerge.

Providers usually fuse data (e.g. geometry, water quantity and quality, air quality, sources of pollutants) and constantly run models to derive current state and forecasts of a particular condition, such as air or bathing water quality risk indicators, to which they associate warnings (e.g. flags or traffic light concept). This qualitative way of showing data may be perceived as not properly substantiated and just as an "opinion". However, the input data that is used to provide this kind of results will have always been validated, as all the processes behind it, and the warnings will reflect stipulated values in according legislation.

This is especially important for the public users, because they are mainly interested in simple and straightforward information, such as the above mentioned environmental quality indicators, and not in the whole process of validation and estimation. This user may be interested to know if the bathing waters are proper for swimming or if the air is polluted, for example. So, he must be assured that these quality indicators are being evaluated according to the limits set by legislation or other applicable regulations.

7. References

- [1] Archetti, F. Messina, E., Toscani, D. and Frigerio, M. (2008) “*KOINOS: Knowledge from Observations and Inference in Networks of Sensors*”, In Proc. SN2008, IASTED International Conference on Sensor Networks, Crete, Greece, 2008.
- [2] Archetti, F., Messina, E., Toscani D. and Frigerio, M. (2008) “*IKNOS – Inference and Knowledge in Networks of Sensors*”. *Submitted to International Journal of Sensor Networks (IJSNet)*, 2008.
- [3] Directive 95/46/EC of the European Parliament, Journal of the European Communities of 23 November 1995, No L. 281 p. 31.
- [4] Klein, A., Do, H., Hackenbroich, G.; Kamstedt, M., Lehner, W. (2007) "Representing Data Quality for Streaming and Static Data," *Data Engineering Workshop, 2007 IEEE 23rd International Conference on* , vol., no., pp.3-10, 17-20 April 2007
- [5] Tulone D. and Madden, S. (2006) "PAQ: Time series forecasting for approximate query answering in sensor networks", in *Proc. of the 3rd European conference in Wireless Sensor Networks*, 2006, pp. 21-37.
- [6] INAG, Classificação dos Cursos de Água Superficiais de Acordo com as suas Características de Qualidade para Usos Múltiplos (Water Quality Limits to Superficial Waters With Multiple Uses)
http://snirh.inag.pt/snirh/dados_sintese/qualidadeag/classificacao.html