



**HAL**  
open science

## Managing Sensor Data Uncertainty: a data quality approach

Claudia Gutierrez Rodriguez, Sylvie Servigne

► **To cite this version:**

Claudia Gutierrez Rodriguez, Sylvie Servigne. Managing Sensor Data Uncertainty: a data quality approach. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, 2013, 1, 4, pp.35-54. 10.4018/jaeis.2013010104 . hal-01339140

**HAL Id: hal-01339140**

**<https://hal.science/hal-01339140>**

Submitted on 29 Jun 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Managing Sensor Data Uncertainty: a data quality approach

Claudia C. Gutiérrez Rodríguez (I3S Laboratory, Nice Sophia-Antipolis University, Sophia-Antipolis, France) and Sylvie Servigne (LIRIS – UMR 5205, INSA de Lyon, Université de Lyon, Villeurbanne, France)

## ABSTRACT

With an increasingly technological improvement, sensors infrastructure actually supports many current and promising environmental applications. Environmental Monitoring Systems built on such sensors removes geographical, temporal and other restraints while increasing both the coverage and the quality of real world understanding. However, a main issue for such applications is the uncertainty of data coming from sensors, which may impact experts' decisions. In this paper, we address this problem with an approach dedicated to provide environmental monitoring applications and users with data quality information.

**Keywords:** Data Quality, Sensor Data, Metadata, Environmental Information Systems, Decision Making Support

## INTRODUCTION

Technological improvement on sensors and sensor networks has opened many opportunities to use and combine geospatial data coming from sensors. Today it is much easier to use geospatial sensor data in all sorts of applications as for animal tracking, environmental monitoring, farmlands monitoring, etc. Despite a strong interest from geospatial domain, the integration of sensors raises new technical and data management challenges, like sensor data uncertainty. In fact, sensor devices have hardware restrictions and perform data collection in hostile environments turning data more imprecise and uncertain. Moreover, the quality of sensor data is often decreased by sensor failures or malfunctions. Thus, deficiencies on sensor data cannot be ignored, but tackled in order to reduce information misunderstanding and assist experts in decision making process.

Data quality is considered of high importance in the management of all Information Systems (IS). Thanks to several research and standardization programs (i.e. Devillers & Jeansoulin, 2006; ISO19113, 2002); Geographic Information Systems (GIS) have a very complete data quality management strategy dealing with data uncertainty. Recently, international organizations in this domain as Open Geospatial Consortium (OGC), the National Oceanic and Atmospheric Administration (NOAA) and the National Aeronautics and Space Administration (NASA), have also raised the importance of analyze the quality of geospatial sensor data and strongly advice to perform further studies.

In this paper, we introduce an approach attempting to manage the uncertainty of sensor data according to data quality principles, especially for environmental monitoring applications. Based on sensor data specificities in this particular context and inspired by current methodologies and approaches for manage Data Quality (DQ) and Quality of Service (QoS), our approach proposes a data quality model able to formalize the characteristics, requirements and constraints of sensor data. Besides, we propose a sensor data quality processing which allows us to provide a quantitative and qualitative representation of sensor data quality. And finally, inspired by visualization approaches in GIS, we propose to communicate via an user interface information concerning the quality of sensor data. We exploit the visual representation of quality indicators and the generation of data quality reports.

The rest of this document is organized as follows: the first section introduces a discussion of our research context and challenges. Next, we describe the related work motivating our approach. The third section is dedicated to introduce our proposal of sensor data quality definition and modeling. Forth section attempts to describe our approach to evaluate the quality of sensor data. In the fifth section, we detail how we propose to communicate quality information to users and applications. Then, sixth section describes our proposal of visualization interface for sensor data quality discovery in a monitoring context. Our conclusions are detailed in the last section.

## **THE PROBLEM OF SENSOR DATA QUALITY**

For years, data quality characterizes a key problem for all kind of organizations (Wang & Strong, 1996). Actually, emerging applications in geospatial domain and manipulating sensor data also reveal this important issue. Considering monitoring as a primary key on environmental crisis management systems, an early and reliable detection of critical events is crucial for systems achievement. Environmental monitoring thus requires an efficient acquisition of information coming from sensors (spread over large areas), an interpretation of complex observation patterns at different temporal and spatial scales, as well as reliable and understandable results. These facts led us to wonder: *how to evaluate and provide users with quality information?*

### **Case of Study: Environmental Monitoring Applications**

Our research is motivated by the analysis of data quality especially for environmental monitoring applications. With our approach, we aim to identify and tackle the most transcendental aspects of this problematic.

An Environmental Monitoring System (EMS) refers to the activities or processes used to characterize and monitor the environment (Maier & Vanstone, 2005). Generally, this kind of systems is used to support environmental risk management and evaluate its impact. Through the years, the EMS have become a key of monitoring programs all over the world and applied to control a variety of chemical, biological, or radiological aspects for instance (Environment Agency, 2005).

Several sampling methods and techniques are actually employed in environmental monitoring such as remote sampling and sensing or continuous monitoring (Karabork, 2010; Briem, Benediktsson, & Sveinsson, 2002). We are particular interested in those infrastructures deployed to perform a remote environmental monitoring, especially in a continuous manner, such as

natural phenomena monitoring systems. This kind of systems is composed by a set of wireless, geolocalised and heterogeneous sensors, also called *geosensors* (Trigoni, Markham, & Nawaz, 2009). Such sensors are typically organized in networks, measuring one or several parameters (i.e. temperature, movement,...) and especially deployed in contexts where an environmental activity carry a risk of harmful effects both for human safety and/or for the natural environment.

In Figure 1 we depict a typical framework for such a system. Here, an environmental phenomenon or activity is continuously, periodical and remotely monitored by sensors.

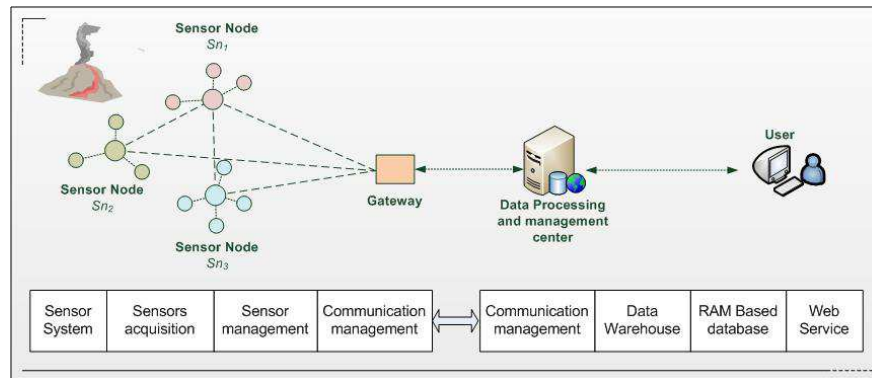


Figure 1. Typical framework of a natural phenomena monitoring system.

Sensed data are collected and pre-processed to be transferred (directly or through a gateway) to the data processing and management center for further analysis. Collected data are processed, managed and stored according to users and application requirements. Furthermore, such data can be also integrated with other information such as satellite images or maps. Resulting data can be visualized and additional analyzed by a GIS or published and analyzed through the Web by specialized services (i.e. (OGC, 2007)). This kind of framework is also employed to monitor farmlands, pollution, animals or smart buildings, among others (Trigoni, Markham, & Nawaz, 2009).

### What about sensor data uncertainty?

Considering such EMS specificities, we observe several important aspects producing an impact on the quality of sensor data. First, such monitoring systems impose sensors to operate in difficult conditions. In addition, sensors perform measurements according to their restricted storage, processing and communication capabilities. In both cases, sensors can provide unreliable information. This effect is very critical because, an erroneous data can be propagated and provide a misinterpretation of the real world.

However, we note that a given data may not only be erroneous because of a faulty sensor; deployment conditions of the sensor and their use have also an impact (Ni, Ramanathan, & Nabil, 2009). Generally, erroneous data coming from sensors can be identified in two categories: *intentional* or *unintentional* errors (Ni, Ramanathan, & Nabil, 2009; Shi & Perrig, 2004). Intentional data errors can be caused by physical or logical attacks over sensors (i.e. malicious attacks) and unintentional data errors are generally caused by hardware malfunction (or stochastic errors), misplacement (conditional errors) or exhausted resources (systematic errors).

Regarding unintentional errors, sensors used to monitor environmental phenomena are exposed to unpredictable situations. For example, a sensor placed in the perimeter of a mountain can collect measurements with temperatures outside of the specified patterns. At first glance, this leads us to conclude that there is something wrong with the sensor. But this effect can be avoided if we take into account that such a sensor is deployed in an area where a snow storm struck few hours ago, and the sensor is now two feet below the snow. The sensor is working properly indeed, but environmental conditions are not those predicted and strongly affect the attended measurements. Multiple additional problems can also be derived from or caused by animals surrounding the detection area, for instance.

Considering sensor data quality issues and facts, we argue that monitoring systems based on data coming from sensors and sensor networks must be reliable enough to guarantee system achievement and assist experts on decision making. While no system can actually guarantee data quality, our concern is how to provide users and applications in a monitoring context with quality information. We address this topic area with an approach attempting to define, evaluate and communicate the quality of sensor data.

Hereafter, we introduce in more detail the existing literature inspiring our work and being helpful in understanding the scope of sensor data quality issue.

## **RELATED WORK**

According to the literature, data quality issue has been the subject of numerous studies, especially for Information Systems (Naumann & Roth, 2009). Through literature, the research community has tried to explain the meaning of data quality; a widely-accepted description is “*quality data are data that are fit-for-use by the data consumer*” (Wang and Strong, 1996). However, the diversity of data environments leads to numerous approaches dealing with this issue at various application domains (i.e. geographic, medical, military...) (Batini, Cappiello, Francalanci, & Maurino, 2009). Reviewing these approaches, we note that data quality is characterized differently according to: *the type of IS* (datawarehouse, distributed...), *the specificities of data* (type and variability: xml, dynamic) *and their processing level* (raw, components, products...) (Wang & Strong, 1996; Strong, 1997; Pipino, 2002; Scannapieco, 2004; Peralta, 2004; Naumann & Roth, 2009). For example, the analysis of data quality over systems processing homogenous and static data remains simpler than for those distributed systems processing heterogeneous and multisource data. Traditionally, the interest of DQ research is then focus on the definition of *dimensions* (i.e. accuracy, consistency,...), *models* (i.e. Multidimensional Spatial Data Quality Model), *methodologies* (such as AIMQ – A Methodology for Information Quality Assessment), *techniques* (i.e. metadata, data mining...) (Batini, Cappiello, Francalanci, & Maurino, 2009) and tools adapted to all sort of IS (Peralta, 2004).

## **Data Quality and GIS**

At the present time, Geographic Information Systems (GIS) have a very comprehensive strategy for data quality management. As (Devillers, Stein, Bédard, Chrisman, Fisher, & Shi, 2010) explains, geographic research community increasingly focuses on spatial data uncertainty or spatial data quality, since 1980's. Evaluation and improvement approaches (Goodchild, 1998; Hunter, 2001; Devillers, Stein, Bédard, Chrisman, Fisher, & Shi, 2010; Lassoued,

Bouadjenek, Boucelma, Lemos, & Bouzeghoub, 2010) as well as standardizations programs are the basis of geospatial data management (Jakobsson & Giversen, 2007).

In this particular domain, any data source is not limited only to attribute data; it is fulfilled with information characterizing the source itself as *metadata* (Hunter, 1999; Devillers R. B., 2002). As a result, data quality is modeled according to two main trends, one oriented to *intrinsic* or *internal* characteristics of data typically included in metadata (acquisition properties, modeling, storage...) and a second one referring costumers' satisfaction with *fitness-for-use* or *external* characteristics (Devillers & Jeansoulin, 2006). These two trends lead to a set of data quality characteristics, actually suggested by research community and standardized by five or seven criteria according to the application and requirements (ISO19113, 2002), such as: *lineage*, *spatial/positional accuracy*, *attribute accuracy*, etc. Techniques and mechanisms to evaluate the quality of spatial data are also standardized (ISO19114, 2003) and supported by various representation techniques (i.e. cartography symbols, reports...) which are used to express data uncertainty in an intuitive and effective manner (Hunter, 1999; Devillers R. B., 2007; Huth, 2007).

In order to monitor the environment, the geospatial domain has also shown a continuous improvement with the use and discovery of geosensors, especially with the Sensor Web Enablement initiative (SWE) (OGC, 2007). This initiative enables an interoperable discovery of sensor resources in a standardized way (OGC, 2012; SSN Sensor W3C, 2012) but also has raised important data quality concerns.

### **Quality of Service (QoS)**

Actually, research work focuses more and more on improving sensors infrastructure and behavior. Increasingly, proposals to qualify and maintain sensor resources are proposed. We distinguish for example approaches to analyze sensor as a service (QoS) (ISO/IEC13243, 1999) and qualify them as such (Ahluwalia & Varshney, 2009); others propose filtering, cleaning or clustering techniques to limit faulty data propagation (Ni, Ramanathan, & Nabil, 2009), etc.

Our perception of quality over a sensor is strongly related to the Quality of Service (QoS). Formally, QoS is seen as “a *set of characteristics related to a collective behavior of one or more objects, in order to determine the utility of a service in a specific application context*” (CISCO, 2001; ISO/IEC13236, 1998; ISO/IEC13243, 1999). In Information Technology (IT) domain, QoS is modeled by several elements associated to the functional and non-functional aspects of a service as: *categories, characteristics, rules, policies, management functions and mechanisms* enabling users to prescribe particular *QoS requirements* (Figure 2). According to this model, some general QoS categories are for example those related to *time, consistency, integrity, reliability*, etc., and each category includes one or more characteristics used to represent one or several aspects of a service or of an identifiable and quantifiable information resource (i.e. device, database, etc.). Functions and mechanisms are applied following specified management activities like *QoS monitoring, QoS maintenance or regulation*, etc. (ISO/IEC13236, 1998).

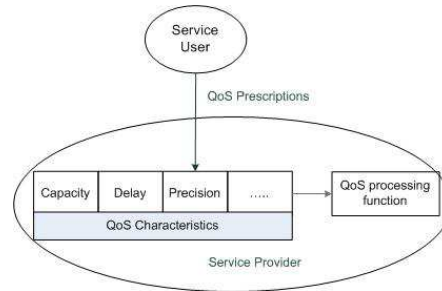


Figure 2. Terminology of (QoS) (ISO/IEC13243, 1999)

Quality of Service is also a notion actually used in GIS, especially for those approaches using geographic information services or geo-services (Onchaga, 2004; ESRI, 2010). Here, QoS comprises desirable qualities on geographic information delivered by a chain of geo-services and the qualities associated with the collective behavior of the geo-services (and other services) that create the service chain. Actually, geo-services infrastructure allows to discovery relate and execute geo-services according to user's needs (ESRI, 2010; Simonis, 2005).

Through the literature, we can observe that current approaches provide excellent quality and data quality principles enhancing the traditional perception of data quality, and motivate the spread of quality assurance to emerging applications. However, these approaches are mainly focused on so-called traditional applications, where static data is managed in databases management systems or datawarehouses. Regarding sensor data quality requirements in a geospatial context, such approaches do not sufficiently take into account the properties of geosensor data, namely their *dynamicity*, *temporality* and *heterogeneity*. As a result, the necessity to tackle the quality of sensor data over monitoring systems remains a relevant research topic.

## DEFINING SENSOR DATA QUALITY

As we observe, data quality is not an easily definable term, it has many different facets and its meaning varies across different aspects such as requirements, users, etc (Redman, 2001). Each individual perception of quality will vary, especially depending upon their context. Accordingly, our approach attempts to define sensor data quality following two main criteria: the *sensor data specificities* in the monitoring context and their *quality requirements*.

The rest of the section introduces in more detail these aspects, especially concerning the review and modeling of sensor data specificities exploited in environmental monitoring applications and our proposition of sensor data quality modeling.

### Sensor Data Specificities

In a monitoring context, sensor data has several specificities comparing to data exploited in traditional applications. To better explain this aspect, we propose to formalize sensor data specificities with a product-based point of view: from *acquisition* to *utilization*.

In order to reach this goal, we decouple the typical framework of environmental monitoring systems (Figure 3) in three main layers: *acquisition*, *processing* and *utilization*. Such a decoupling allows us to abstract sensor data specificities all through the system.

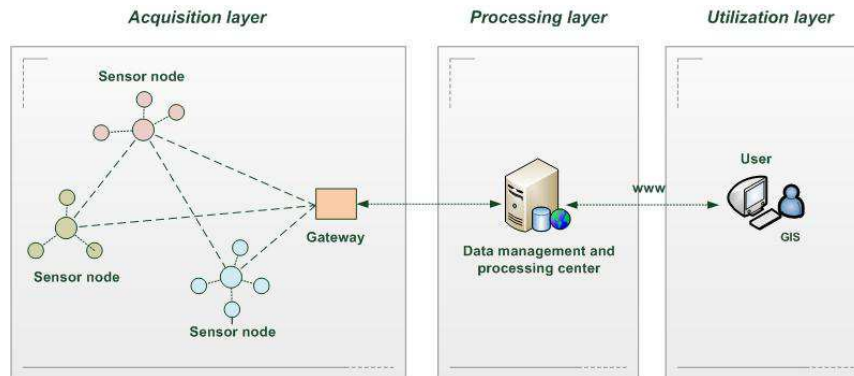


Figure 3 – Decoupling a typical EMS framework.

The *acquisition layer* refers to the sensor data collection system where sensors, raw (or sensed) and pre-processed data are managed. The *processing layer* involves data resulting from data processing and management center where energy, storage and analyze capabilities are more significant. Finally, at *utilization layer*, we talk about delivered data (or post-processed data) exploited over a GIS or combined with Web services or applications.

According to these layers, we distinguish several specificities. For example, data coming from sensors are geolocalised and time stamped values. Sensor data have then spatiotemporal properties and mainly stored over temporal and spatial relations. Also, for environmental monitoring systems, “*real-time*” processing does not means “*fast*”, it means that processing time is limited, predictable and manageable, also called *soft real-time*. In a complementary manner, to process all sensor data, monitoring systems must adopt priority policies as preferring most recent data and store them according to temporality properties. Finally, sensor data can be further processed by users with a GIS or a complementary geo-service producing statistics, sorting or geo-locating data in a map, for example.

We consider that the spatial, temporal and dynamic properties of data coming from geosensors introduce a new scheme of data collection. Since *observation* is the principal goal of monitoring systems, we base our data modeling on this concept (also used by OGC – SWE (OGC, 2007)). Thus, we call *observation* data those data used to describe a phenomenon owning spatial, temporal, semantic and dynamic properties as well as complementary information contained in *metadata*.

In Figure 4, we attempt to provide a big picture of data coming from sensors and the complementary information (metadata) which characterizes data in a monitoring context. Being compatible with current applications (OGC, 2007; SSN Sensor W3C, 2012; Swiss Experiment, 2010) in geospatial domain, our model consider that a sensor network is composed by a set of sensors, located on the same observation area and allowed to collect and transfer data at fixed and variable positions (fixed, agile and mobile). Such sensors are related to observation stations (meteorological, agricultural stations...) responsible for observing different phenomena (i.e. tsunamis, volcanoes ...), and where one or more elements are used to determine the evolution of such phenomena (i.e. temperature, gas, etc.).





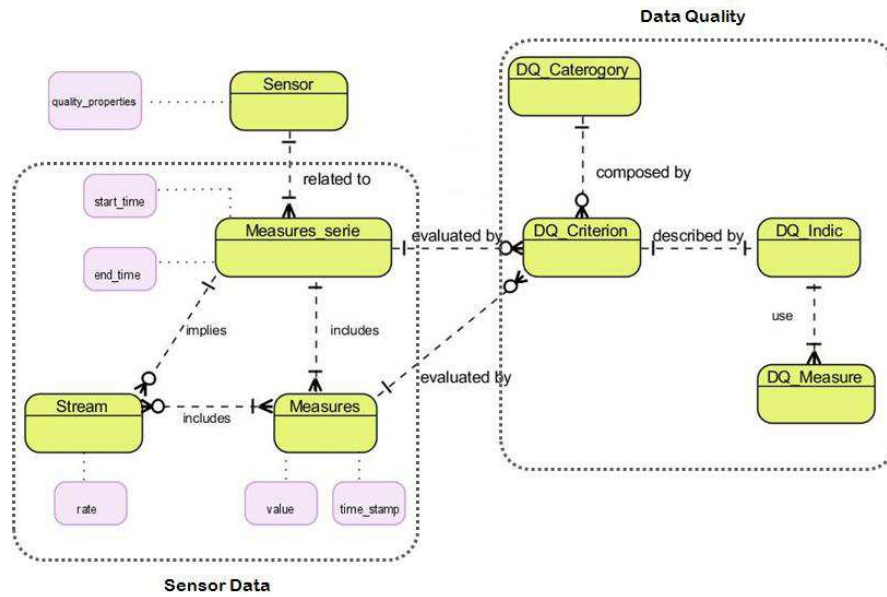


Figure 5 - Sensor Data Quality Model

In Figure 5, we depicted the correlation of our quality principles and the sensor data modeling. Here, quality information is related to sensor data at different granularity levels (a measure and measures series) and taking into consideration the sensor quality properties. Thus, one or several criteria can be related to a particular measure, to a set of measures or to a stream of measures from a given sensor. Evaluation results can thus be depicted by quality indicators. More specifically, a *data quality category* refers to a quality component within a monitoring system (i.e. context, processing and discovery). A *data quality criterion* can be considered as an extension of data (qualitative or quantitative) and referring to factors impacting quality as reliability, accuracy, etc.

### Sensor Data Quality Criteria

In order to categorize the set of quality elements necessities to define and estimate the quality of sensor data, we expose first the set of factors that can affect data with a certain degree of uncertainty and thus impacting its quality. Subsequently, we will associate these factors to the quality principles suggested in our sensor data quality model. This process allows the impact estimation.

As we state in before, sensor data is processed at different levels of the monitoring system. At each level, the quality of sensor data can be impacted by several factors and thus, erroneous or poor quality of data can reach the final user. To determine these factors, we analyze each layer of the system as follows.

At acquisition layer, we study the impact factors according to three main aspects: *measurement context, sensor and transmission*. Here, we identify several factors or properties such as sensor calibration and performance, battery level, storage and processing capacity, measurement rate, accuracy and precision, transmission rate and type. Secondly, at processing layer the quality of data strongly depends on the processing mechanisms used to transform data

as: raw *sensor data gathering and validation, data processing and storage*. More explicitly, factors like processing mechanisms (e.g. aggregation), the presence or absence of data validation mechanisms (e.g. filtering), the quality of service of the main server (storage level, availability, server load...) as well as the processing time and temporality of sensor data (i.e. update, historical, recent). Finally, at the discovery layer quality factors are related to *how data is extracted, represented and queried*. In this instance, we identify factors like automatic extraction mechanisms, representation models or the human factor.

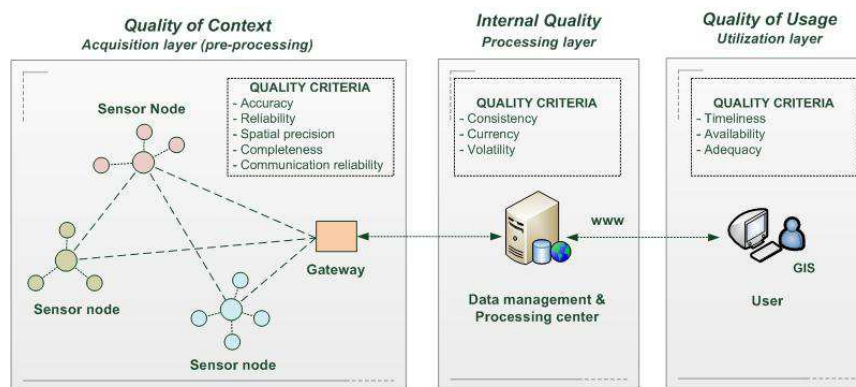


Figure 6. Data Quality Criteria for Sensor Data

Considering the nature of impact factors, we estimate that sensor data quality can be expressed as “*sensor data complying the reliability, temporality and adequacy of data used for an intended goal*”. And as a result, we recommend several quality criteria adapted to our approach and essential for the evaluation of sensor data quality. We illustrate in Figure 6 the results of this study. We classify then quality criteria into three categories: *context, internal and usage*.

The first category regroups the set of criteria selected to estimate the quality of raw sensor data at the acquisition layer: *accuracy* (correctness data according to a reference value and sensor technical precision), *reliability* (estimated value considering a set of possible random sources over a data source and a probability of reliability), *spatial precision* (comparison between an initial or current spatial reference with an estimated value regarding the precision factor of positional system), *completeness* (comparison between an estimated quantity of produced data for a given time and rate and the current values), *communication reliability* (estimated value considering signal strength, data package and a theoretical noise factor provided by technical specifications). Such criteria allow us to estimate the quality on data sources, their context of acquisition and their transmission to the data management and processing center.

The *internal category* is related to quality criteria such as *consistency* (comparing acquired and expected data), *currency* (degree to which information is current or updated) *volatility* (value representing the variation of data over time). Their main goal is to avoid inconsistent information and to maintain the temporality of sensor data at a processing level.

Finally, *usage category* includes criteria such as *timeliness* (measure representing the comparison between the time delay at which data is available and the time when data availability

advertising appears), *availability* (measure representing the accessibility of data for an intended use) of and *adequacy* (estimation of usability or quality of use (QoU) (ISO/IEC25010, 2011)).

## **SENSOR DATA QUALITY ASSESSMENT**

In our approach, we are committed to provide users and applications with information about the quality of sensor data in a monitoring context (real-time and differed-time). In our study, we observe that the main interest from experts the environmental monitoring field is to have means to make decisions as most reliable as possible. Hence, we focused on qualify and communicate quality information instead of a correction or improvement of sensor data. In order to reach this goal, we have to accomplish three main tasks: *specify the information quality sources, estimate sensor data quality and manage data quality information.*

In this section, we introduce first a proposal for defining and structuring quality information sources. Such structure is based on current metadata standardizations in geospatial domain and modeled according to the specificities of sensor data in a monitoring context. Next, we propose to estimate the quality of sensor data based on a multicriteria approach based on sensor data quality model. Finally, we describe our initiative to manage quality information.

### **Data quality sources: Sensor Data and Metadata**

The assessment of sensor data quality implies a strong correlation between sensor data and the information about dynamic changes of quality values. Sensor data are traditionally enhanced with contextual or complementary information like sensor battery level, position, etc. which are thus structured in metadata. In our approach, we consider that quality information is also complementary information describing the uncertainty of sensor data and helpful to sensor data understanding, and thus metadata are data related to sensors behavior, the specificities of monitoring context and to data quality information. In order to manage such information, we propose a metadata modeling oriented to structure complementary information. This model is inspired on current metadata standardizations in geospatial domain and respects the *dynamicity* (time and space), *granularity* (abstraction level) and *generality* (generic or applicative) of sensor data.

In Figure 7, we depict sensor metadata structuring considering: *Observation, Sensor and Quality.* Firstly, *Observation metadata* refer to information describing the specificities and goals of an observation (cf. Sensor data modeling). For this kind of metadata, complementary information is recognized as general (i.e. observation type, id, temporality, etc.) or referring to observation's responsible part (Responsible name, organization, email, etc.). Secondly, *Sensor metadata* refer to both static and dynamic information about a sensor, allowing us to identify and evaluate the capacities of a sensor at an instant of time. On the one hand, *Static sensor metadata* contain the basic information to identify a deployed sensor considering four components: *Sensor general information, Sensor features, Sensor operational constraints and Sensor interventions.* On the other hand, *Dynamic sensor metadata* refer to information required to verify sensor's status all through an observation.

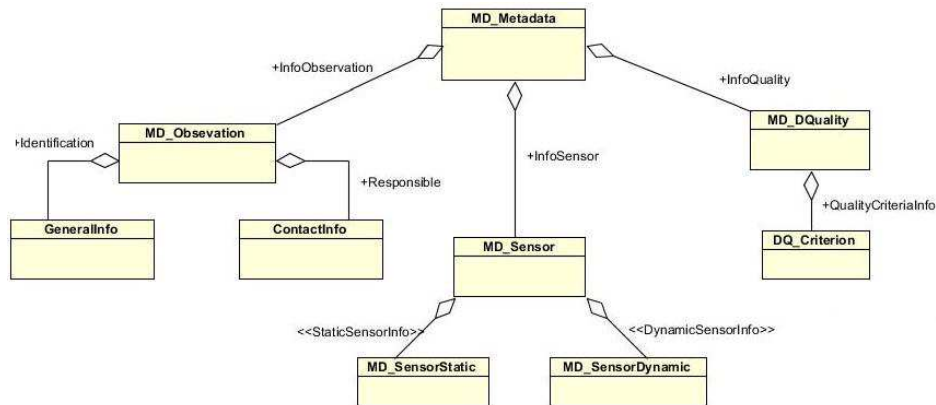


Figure 7. General Metadata Model

These metadata contain information as: *Sensor status, sensor battery level, Sample rate, Storage level, Transmission rate, Mobility rate, Sensor coordinates*. Finally, *Data quality metadata* manage information indicating the quality properties of data that we are evaluating such as *criteria, measure, indicator* etc. Such metadata includes the description of quality assessment principles previously described (cf. Data Quality Model for Sensor Data).

If we consider that in an observation, a sensor or sensor node provides information related to measured values and metadata. Measures are then sensed values (temperature, pressure...) and metadata refers to information like sensor lineage, behavior and quality. Consequently, each object resulting from a data, a dataset or a datastream is related to quality principles and communicated to the user by the means of quality indicators or data quality reports. However, unlike traditional metadata management, metadata in a monitoring context is characterized by the real time processing and the dynamicity of information. As a result, good policies for metadata extraction and update are required.

We consider that two kinds of metadata extractions in real-time are possible: *automatic* or *on-demand (user)*. Indeed, metadata are generally used in two main cases: one to access targeted information in response to occasional user queries (*on-demand*) or to provide information previously specified (*automatic*). We estimate that an automatic extraction can be performed over static information, according to application and user requirements and respecting system's constraints. Metadata over dynamic information can also be provided automatically, but in order to avoid system overload, the temporality of such information will be reduced. Dynamic or more specific information (those not imperative to understand the behavior of an observation) is suggested to be extracted on-demand by the user at instant or period of time.

Once metadata extraction is performed, the metadata update process implies also several techniques allowing user to discovery current sensor information (i.e. trigger, manual, automatic). Contrary to static metadata, complementary information with high rate of variability must stay updated in order to capture temporal changes in information and thus, metadata management becomes more difficult in these circumstances. As a result, we propose to update metadata according to two techniques: *periodic* (by time windows), *trigger* (occasional or by

detection). Such technique allows economizing system's resources, coordinating the system and avoiding information overload.

### **Processing Sensor Data Quality**

According to our sensor data quality model and quality information sources structuring, we are able to provide user with information qualifying data that he/she uses. Our approach proposes a multicriteria evaluation of sensor data quality based on (Berti, 1999). In such an evaluation, one or more quality criteria can be assigned and evaluated to represent the uncertainty or reliability of sensor data. It can be implemented at different points of the system's framework, especially at before data discovery. Thus, evaluation results can be coupled to sensor data and being reused and discovered to at the same time.

In order to estimate the uncertainty or quality level of a data, a dataset or a datastream referred as objects, we first consider that an instant or period of time, for each object has one or several attributes  $k_{i=1 \rightarrow n}$  associated to measurement values  $v_i$  used to estimate each quality criteria  $Q_c(k)$ . Thus, for each object a set of quality values representing quality criteria will be included. The number of quality criteria for each object is not fixed and can be adapted according to its specificities (type of object, variability, etc.). This perception allows also an extension of the processing mechanism to other quality criteria. Finally, a global vision of a given object  $Q_\omega(m_i)$  can be done by the product of a weight factor  $W_k$  for each estimated quality criteria  $Q_c(k)_i$ .

$$Q_\omega(m_i) = \sum_{k=1}^n W_k Q_c(k)_i$$

In order to allow an acceptable data quality information discovery, we propose to structure sensor data with quality information and thus quality results. In this way, quality information can be used as a reference for further analysis and support experts in their decisions.

Next section is dedicated to describe how we estimate to provide experts and applications with quality information.

### **COMMUNICATING SENSOR DATA QUALITY INFORMATION**

In the geospatial domain, several approaches have been focused on finding better ways to communicate data quality and assist experts. These proposals are mainly based on procedural methods (Hunter, 2001) on visualization techniques (Devillers R. B., 2007; Huth, 2007; OGC, 2007), on visual or audio signals to alert users (Gervais, 2006) as well as on data quality reports (ISO19114, 2003). In general, these approaches attempt to communicate the possible errors in terms of quality in a geographic area, or using information provided by quality criteria to inform the user the status of a geometric component, a data or a dataset, etc.

### **Sensor data quality monitoring and reporting**

Unlike to traditional applications in geospatial domain, our research scope is confronted to static and dynamic geospatial information describing the behavior of an observed phenomenon and associated to the quality of sensor data, among others. Hence, adapted methods to visualize

such variability are required, especially avoiding information overload or misunderstanding which can easily confuse the user.

Mainly inspired by (Devillers R. B., 2007; Huth, 2007; OGC, 2007; Hunter, 2001), we chose to characterize sensor data quality information by associating metadata with a visual representation of quality indicators and generating sensor data quality reports. We estimate that regrouping such quality information sources, we are able to provide a more dynamic and forthcoming quality information discovery. In Figure 8, we illustrate several suggested representation methods that we study and which summarizes the compatibility of such methods with the specificities of sensor data quality, especially considering dynamicity and temporality.














	Suitable	Memorable	Adding text	Understandable	Accurate
 Colored rectangle	++	++	++	++	--
 Bar	++	++	++	++	++
 Slider	++	++	+-	++	++
 Circle segment	--	--	+-	+-	--
 Traffic lights	++	++	++	++	--
 Variable graph	+	+-	++	+-	--
 Line	--	--	++	--	--
 Narrow	--	--	++	+-	--
 Plus/minus	++	++	+-	++	--
 Circle	--	+-	++	++	++
 Counter	++	++	+-	++	++
 Face	++	++	--	++	--
 Digital	++	++	+-	+-	++

Figure 8. Comparative table of visual methods and suitability criteria for representing the quality of sensor data (Devillers R. B., 2007; Huth, 2007)

Together and according to geospatial research community, quality reports are also very important for both information users and producers (Jakobsson & Giversen, 2007). In fact, a report of data quality assessment is essential to communicate, analyze and manage data quality *a posteriori*. Specially based on the principles of ISO 19114 standard and our metadata management approach, we determinate the information to be included in this report. By definition, several elements must be included in this report such as: *report id, date and analyzed scope, quality criteria, measures, indicators, evaluation methods, etc* (ISO19114, 2003).

In our approach, a report can be generated depending on user preferences or according to system's behavior. For example, we can generate a report at each observation ending or when user requires (at the beginning, 30 minutes before the end, when some quality discrepancies appear...). In some other cases, when data quality is not an active concern, a quality report is not frequently required. In order the dynamicity of the environment and user requirements, we propose a process for sensor data quality report (Figure 9). In this process, we start to select the scope to be observed or analyzed and the set of quality indicators to be gathered. Then, the corresponding contextual information to both selections will be aggregated automatically. Next step remains to program the periodicity of report generation, an instant report generation or a

periodical. In order to maintain as much as possible the consistency between data and quality reports, we decide store quality report independently of data.

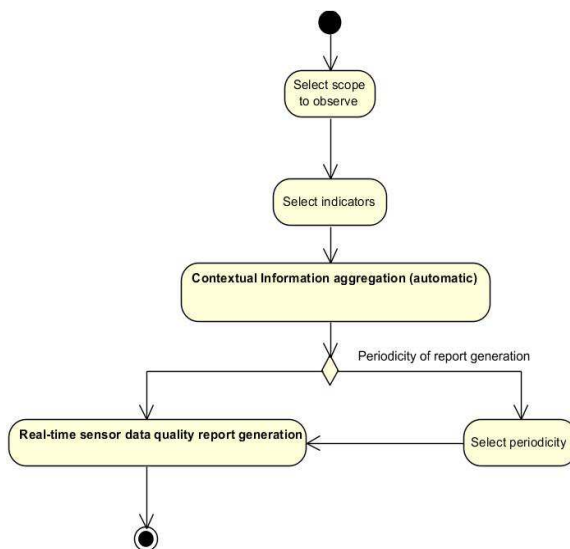


Figure 9. Process for Sensor data quality report

We recognize that using this generation process, a minimal risk in terms of timeliness of information can arise, especially because data updating is done once the report is generated. However, this process is suitable to prevent non-required reports generation.

## SENSOR DATA QUALITY DISCOVERY

In order to assist experts in the assessment of sensor data quality, we propose a prototype of visualization interface. Our prototype allows the visualization and discovery of data coming from sensors together with the communication of contextual information such as monitoring information and the quality of sensor data. This prototype allows us to validate our approach and implement several mechanisms as geospatial data discovery, sensor data quality management, as well as the use of visual quality indicators and reports. This prototype proposes users a way to interact with sensor data and quality information in a monitoring system.

### Scenario: monitoring volcano activity

We take as an example the surveillance of a Mexican volcano: *Popocatepetl*. The Popocatepetl is localized at 60 kilometers from Mexico City and it extends on three Mexican states, where each one of them has its own disaster management policy to protect population, farmlands and industry developed in the surroundings. This volcano is under the surveillance of the National Center for Disaster and Prevention of Mexico (Cenapred) which employs a set of sensors distributed on 25 stations and process approximately 64 telemetry signals with 16 computers. Such monitoring system supervises in a visual, seismic, geodetic and geochemical way the behavior of the volcano. As this case of study shows, users of environmental monitoring systems are interested on discovering information about the evolution of the observed phenomena.



## MoSDaQ prototype

The MoSDaQ (Monitoring Sensor Data Quality) prototype refers to a web-oriented user interface intending the monitoring of a natural phenomenon. This prototype attempts to be exploited through the Web by experts (locally or remotely) at a client side (Figure 10).



Figure 10. MoSDaQ a Web-oriented user interface.

This interface is composed by five main sections: *mapping*, *observation*, *sensor information*, *data querying* and *quality*. The *mapping* section refers to the localization of sensor objects in the cartographic space, what we call observation zone.



Figure 11. MoSDaQ – Mapping section

In this zone, we place several sensors in a given position (by coordinates) and characterized by an icon according to its type and status. By clicking on each sensor icon (Figure 11), we have access to sensor information (*static and dynamic*).

*Observation section* introduces all the information related to observed phenomena and elements. This section notify four important aspects: *observed phenomenon*, *observation features*, *measured elements* and *coordinates of observation zone*. Also, *query section* allows querying current or historical observation data according to its spatial, temporal, quality or semantic properties. *Sensor description* section provides all technical information about each sensor located in the observation zone, as type, supplier, operational features and constraints (Figure 12).

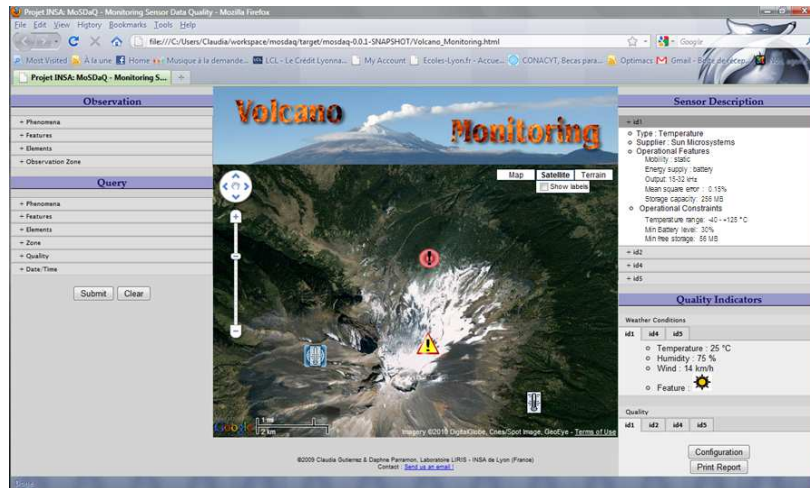


Figure 12. MoSDaQ – Sensor Description

Besides, *quality section* enables the access to quality properties of sensor data according to each deployed sensor in the observation zone. To setting and visualize them, we designed this section with a dashboard using suggested representation methods for quality indicators.

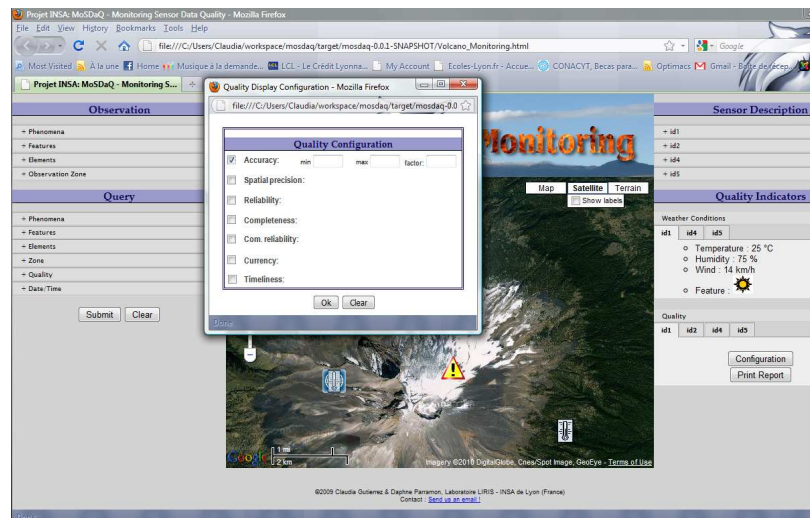


Figure 13. MoSDaQ – Quality Criteria configuration

In a first attempt to interact with user preferences, we propose a customize window, allowing user to setting its quality constraints for each selected criteria (Figure 13). In this section, we can also produce and visualize a sensor data quality report (Figure 14).

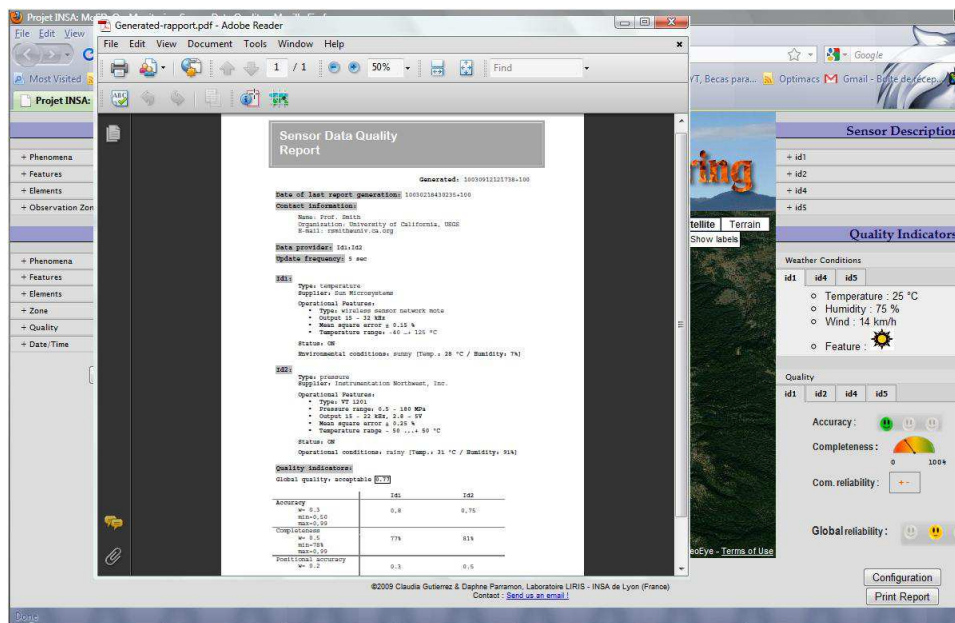


Figure 14. MoSDaQ – Sensor Data Quality Indicators and Report

With this prototype we made a first attempt to demonstrate the feasibility of discovering sensor data together with complementary information as quality information in a monitoring context. This prototype enabled us to explore how to communicate experts' requirements and how to visualize data quality in a dynamic context. In fact, based on primary results, further prototype improvements are already considered. For instance, migrate to a more dynamic and large scale context, exploiting new Web 2.0 development facilities and interact with HCI (Human Computer Interaction) experts in order to propose better ways to interact with users, etc.

## CONCLUSION

In this paper, we underline the importance of evaluating and communicating information quality in emerging geospatial applications like environmental monitoring. As our research work shows, several external factors may impact the quality of sensor data and thus directly impact users' decision making. Hence, we tackle this issue with an approach focalized on the definition and evaluation of sensor data quality in geospatial monitoring applications. We propose a sensor data quality model based on sensor data specificities and attempting to formalize sensor data quality properties (categories, criteria and indicators). This contribution is mainly supported by mechanisms and techniques as metadata to process and manage the quality of information. We also propose in this paper, how to provide users and applications with quality information : visually representing quality indicators and producing quality reports. In order to support our approach, a web-based interface for sensor data quality discovery in real-time has been implemented. This interface is based on the specificities of a volcano monitoring system.

## REFERENCES

- Ahluwalia, P., & Varshney, U. (2009). Composite quality of service and decision making perspectives in wireless networks. *Decision Support Systems* , 46 (2), 542-551.
- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surveys* , 1-52.
- Berti, L. (1999). Quality and Recommendation of Multi-source Data for Assisting Technological Intelligence Applications. *DEXA'99* (pp. 282-291). LNCS 1677, Springer-Verlag.
- Briem, G. J., Benediktsson, J. A., & Sveinsson, J. R. (2002). Multiple classifiers applied to multisource remote sensing data. (I. Transactions, Ed.) *Geoscience and Remote Sensing* , 40 (10), 2291-2299.
- CISCO. (2001). *Internetworking Technology Handbook*. Retrieved 2012, from Cisco: [http://docwiki.cisco.com/wiki/Internetworking\\_Technology\\_Handbook](http://docwiki.cisco.com/wiki/Internetworking_Technology_Handbook)
- Devillers, R. B. (2002). Spatial Data Quality : From Metadata to Quality Indicators and Contextual End-User Manual. *Proceedings of OEEPE-ISPRS Joint Workshop on Spatial Data Quality*.
- Devillers, R. B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science* , 261-282.
- Devillers, R., & Jeansoulin, R. (2006). *Fundamentals of Spatial Data Quality*. London, UK: ISTE.
- Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *T. GIS* , 14 (4), 387-400.
- Environment Agency. (2005). *Environmental Management Systems: Position Statement*. Retrieved from Environment Agency: <http://www.environment-agency.gov.uk/research/library/position/110118.aspx>
- ESRI. (2010). *GeoServices REST Specification*. Retrieved 2012, from ESRI: <http://www.esri.com/industries/landing-pages/geoservices/geoservices.html>
- Gervais, M. B. (2006). Auditing spatial data suitability for specific applications: professional and technological issues. *Workshop on quality assurance in geographic data production*. Marne-la-Vallée, France.

- Goodchild, M., & Jeansoulin, R. (1998). *Data quality in geographic information : from error to uncertainty*. HERMES.
- Hunter, G. (1999). Managing uncertainty in GIS. *Geographical Information Systems* , 633-641.
- Hunter, G. (2001). Spatial Data Quality Revisited. *Proceedings of GeoInfo*, (pp. 1-7). Rio de Janeiro, Brazil.
- Huth, K. M. (2007). Judging and Visualising the Quality of Spatio-temporal Data on the Kakamega - Nandi Forest Area in West Kenya. *5th International Symposium on Spatial Data Quality (ISSDQ2007)*. ITC, Enschede, The Netherlands.
- ISO/IEC13236. (1998). *Information Technology - Quality of Service : Framework*. Geneva, Switzerland.
- ISO/IEC13243. (1999). *Information technology - Quality of service - Guide to methods and mechanisms*. JTC 1/SC 6.
- ISO/IEC25010. (2011). ISO/IEC 25010:2011. *Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models* .
- ISO19113. (2002). ISO 19113: Geographic Information - Quality Principles. *ISO 19113* .
- ISO19114. (2003). *Geographic information - Quality evaluation procedures*. Technical Committee.
- Jakobsson, A., & Giversen, J. (2007). *Guidelines for Implementing the ISO 19100*. eurogeographics.
- Karabork, H. (2010). Selection of appropriate sampling stations in a lake through mapping. *Environmental Monitoring and Assessment* , 163 (1), 27-40.
- Lassoued, Y., Bouadjenek, M. R., Boucelma, O., Lemos, F., & Bouzeghoub, M. (2010). GQBox: geospatial data quality assessment. *GIS 2010*, (pp. 534-535).
- Maier, S., & Vanstone, K. (2005). *Do good environmental*. Bondway, London.: EIRIS.
- Naumann, F., & Roth, M. (2009). Information Quality. In J. Erickson, *Database Technologies: Concepts, Methodologies, Tools and Applications* (pp. 2140-2156). IGI Global.
- Ni, K., Ramanathan, N., & Nabil, M. (2009). Sensor network data fault types. *ACM Trans. Sen. Netw.* , 5 (3), 29.
- OGC. (2012). *OGC Standards*. Retrieved 2012, from Open Geospatial Consortium: <http://www.opengeospatial.org/standards>

OGC. (2007). *The Open Geospatial Consortium, Inc.* Retrieved mars 2010, from SWE - Sensor Web Enablement: <http://www.opengeospatial.org/ogc/markets-technologies/swe>

Onchaga, R. (2004). Quality of service provisioning in geographic information service. *International Journal of Applied Earth Observation and Geo-Information* .

Peralta, V. B. (2004). A framework for analysis of data freshness. *In Proc. of IQIS 2004 (SIGMOD Workshop)*.

Pipino, L. L. (2002). Data Quality Assessment. *Communications ACM* .

Redman, T. (2001). *Data Quality: The field Guide*. Digital Press, Elsevier.

Scannapieco M., V. A. (2004). The DaQuinCIS Architecture: A Platform for Exchanging and Improving Data Quality in Cooperative Information Systems. *Information Systems* , 551-582.

Scannapieco, M. N. (2004). *1st. International ACM SIGMOD Workshop on Information Quality in Information Systems*. ACM.

Shi, E., & Perrig, A. (2004). Designing secure sensor networks. *IEEE Wireless Commun. Mag.*

Simonis, I. S. (2005). Quality of Service in a Global SDI. Kairo, Egypt.

*SSN Sensor W3C*. (2012). Retrieved from Semantic Sensor Network Incubator Group.

Strong, D. K. (1997). *Proceedings of the Second Information Quality Conference MIT*.

Swiss Experiment. (2010). Retrieved Mars 2010, from Swiss Experiment - Interdisciplinary Environmental Research: <http://www.swiss-experiment.ch/index.php/Main:Home>

Trigoni, N., Markham, A., & Nawaz, S. (2009). Proceedings of 3rd Intl. Conference GeoSensor Networks. *GSN2009* (pp. 149-157). UK: LNCS.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems* , 12 (4), 5-33.