Computers & Geosciences 85 (2015) 126–135

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Case study

A quality-aware spatial data warehouse for querying hydroecological data

L. Berrahou [a], N. Lalande [a], E. Serrano [c], G. Molla [a], L. Berti-Équille [c,d], S. Bimonte [b], S. Bringay [e], F. Cernesson [a], C. Grac [f], D. Ienco [a], F. Le Ber [g], M. Teisseire [a,*]

[a] TETIS, IRSTEA, AgroParisTech, F 34000 Montpellier, France
[b] TSCF, IRSTEA, F 63170 Aubière, France
[c] ESPACE DEV, IRD, F 34000 Montpellier, France
[d] QCRI (Qatar Computing Research Institute), Doha, Qatar
[e] LIRMM, Univ. Paul Valery, F 34000 Montpellier, France
[f] LIVE - Université de Strasbourg/ENGEES, CNRS, F 67000 Strasbourg, France
[g] ICUBE - Université de Strasbourg, ENGEES, CNRS, F 67400 Illkirch, France

ARTICLE INFO

Article history:
Received 21 August 2014
Received in revised form
2 July 2015
Accepted 10 September 2015
Available online 24 September 2015

Keywords:
Information system
Data warehouse modeling and design
Data quality
Hydroecological data

ABSTRACT

Addressing data quality issues in information systems remains a challenging task. Many approaches only tackle this issue at the extract, transform and load steps. Here we define a comprehensive method to gain greater insight into data quality characteristics within data warehouse. Our novel architecture was implemented for an hydroecological case study where massive French watercourse sampling data are collected. The method models and makes effective use of spatial, thematic and temporal accuracy, consistency and completeness for multidimensional data in order to offer analysts a "data quality" oriented framework. The results obtained in experiments carried out on the Saône River dataset demonstrated the relevance of our approach.

1. Introduction

Currently, the number of available geo-referenced datasets is increasing with the advent of new spatial data acquisition technologies. These datasets are usually stored and managed using conventional online transactional processing (OLTP) systems, such as spatial database management systems (SDBMS) and geographic information systems (GIS), that combined to spatial analytical tools offer effective decision-making systems. However, these systems present some important limitations when handling huge volumes of data, providing interactive visual analysis and executing aggregation-based decisional queries. In order to handle these issues, data warehouses (DW) and online analytical processing (OLAP) systems have been developed. Indeed as detailed by Kimball (1996) and Bédard et al. (2001), OLTP/GIS systems are optimized for transactional processing (response time and managing concurrency), while OLAP systems are tuned for analytical processing (historical and online analysis). Moreover, contrary to OLTP systems, which handle transactional data (access to hundreds of records), OLAP systems are conceived for analytical data (aggregation and access millions of records). OLAP systems support the decision-making process by allowing users to explore summaries (aggregation) of data to confirm their hypothesis or discover hidden knowledge.

Data warehouses and OLAP systems are widely recognized as decision support systems for the analysis of huge volumes of alphanumeric data modeled using a multidimensional model, which defines the concepts of facts and dimensions (Kimball, 1996). Facts represent the analysis subjects. They are described by numerical measures, which are analyzed at different granularities represented by the hierarchy levels composing the dimensions. The analyst explores warehoused data (cube) through OLAP operators. Integration of spatial data into data warehouses leads to spatial data warehouse (SDW) and spatial OLAP (SOLAP) concepts. SOLAP systems integrate advanced OLAP and GIS functionalities in a unique and effective framework, where the explicit representation of spatial data within dimension levels allows visualization of query results within maps, and the use of topological, metrical and directional operators when slicing multidimensional data (Bimonte and Miquel, 2010). Integration of spatial data as measures enables

* Corresponding author. Fax: +33 467 548 700.
E-mail addresses: lilia.berrahou@teledetection.fr (L. Berrahou), nathalie.lalande@teledetection.fr (N. Lalande), evac.serrano@gmail.com (E. Serrano), guilhem.molla@teledetection.fr (G. Molla), laure.berti@ird.fr, lberti@qf.org.qa (L. Berti-Équille), sandro.bimonte@irstea.fr (S. Bimonte), sandra.bringay@lirmm.fr (S. Bringay), flavie.cernesson@teledetection.fr (F. Cernesson), corinne.grac@engees.unistra.fr (C. Grac), dino.ienco@teledetection.fr (D. Ienco), florence.leber@engees.unistra.fr (F. Le Ber), maguelonne.teisseire@teledetection.fr (M. Teisseire).

http://dx.doi.org/10.1016/j.cageo.2015.09.012
0098-3004/

aggregation of geometric properties, therefore providing a better global understanding of georeferenced phenomena.

For implementing actionable knowledge discovery in Environmental Science, it is worth considering that low data quality will produce unreliable results of statistical analysis and data mining techniques and therefore conduct to misleading conclusions in environmental assessment studies and inappropriate decisions. In our context, we pay special attention to the quality of water quality measurement data because it has an impact on the quality of the results of knowledge discovery from large quantities of water quality measurement data.

Furthermore, ensuring data quality in decision support systems is a hot research topic (Berti-Equille et al., 2011). Indeed, data quality assessment in a multi-source context must be systematically considered, especially in decision-support systems. Current studies have addressed data quality issues within extract, transform and load (ETL) processes by repairing imperfect data, and/or allowing the use of imperfect data in the OLAP system by providing ad hoc conceptual, logical and physical multidimensional data structures and OLAP operators.

Some recent studies have investigated using SOLAP for environmental monitoring (Alexandru et al., 2010; Vernier et al., 2013). They highlighted the interesting analysis capabilities offered by SDW and SOLAP technologies regarding complex environmental data, although they propose a classical SOLAP architecture where data quality problems are handled in the ETL phase, while being transparent in the SOLAP architecture tiers.

This paper presents a new SOLAP systems based approach to deal with complex environmental data, namely hydroecological data, in order to monitor the ecological status of running waters. Here, we refer to SOLAP tools as for kind of SOLAP solutions called OLAP-based (Bédard et al., 2001). In this kind of systems spatial data is only stored in the DW tier, and visualized by means of simply cartographic visualization. This study was carried out as part of the Fresqueau project (2011–2015) which aims at developing new methods for studying, comparing and exploiting all available parameters concerning the status of running waters as well as information describing uses and measures taken. However, due to the size and complexity of the French water network, the data obtained is not of the highest quality and, since sampling costs are very high, this imperfect collected data is the only data available.

Such quality problems lead to unreliable analysis and thus reduce the data warehouse advantages. To cope with these problems, and to ensure the data warehouse efficiency, we propose to integrate data quality dimensions within the warehouse, in order to refine the data analysis. We therefore developed a new method to trace data quality. Our focus on hydroecological data is particularly relevant for characterizing geospatial data quality through various dimensions such as spatial, thematic and temporal accuracy, logical coherence and completeness. Our solution involves a practical approach that combines geospatial data and hydrological parameters' measurements. It is designed to address the complexity of heterogeneous hydroecological data sets. Whereas previous approaches are often overly conceptual, our specific solution has already been implemented in collaboration with hydrologists and continuously optimized. This solution is based on a specific quality data warehouse architecture QuiDA. Data quality concepts and indicators are taken into consideration. This step-by-step decision-making approach offers new data analysis prospects for hydroecologists.

The paper is organized as follows. Section 2 describes the aim of our study and the datasets it is based on. Section 3 is a review on data warehouses and data quality issues. In Section 4, the QuiDA architecture is detailed and illustrated. Experiments on the Saône River dataset are reported in Section 5. The paper ends with a general discussion and some prospects.

2. Motivations and datasets

2.1. The Fresqueau project: studying hydroecological data

The aim of the Fresqueau project (2011–2015), funded by the French National Research Agency (ANR), is to develop new data mining methods for studying, comparing and making effective use of all available parameters concerning the status of running waters as well as information describing uses and measures taken. More precisely, the project addresses two specific issues: (1) gaining further insight into running waters functioning through the analysis of taxons (i.e. aquatic animals or plants) which support biological indices (2) connecting the sources of pressures and the physico-chemical and biological quality of running waters. We therefore rely on physico-chemical and biological data produced by French water agencies in north-eastern (Rhine-Meuse watershed) and south-eastern (Rhône-Méditerranée watershed) France.[1]

According to the European Water Framework Directive (The European Parliament and the Council, 2000), waterbodies are assessed using biological quality elements, based on macro-invertebrates, oligochaeta, fishes, diatoms or macrophytes. Habitat or chemical anthropogenic degradations are also monitored via numerous parameters, generally macropollutants and micropollutants. Therefore, numerous data on the watercourse state are produced yearly from each sampling reach: (i) biological data: faunistic and floristic lists, metrics and indices, (ii) at least six analysis series for each macro-pollutant, (iii) one analysis of different micro-pollutants, and (iv) chemical and ecological states according to the expertise of these results. The data are supplemented by data characterizing the sampling reaches, and data describing the hydrographic network and habitat degradations.

We also collected data estimating human activities (land use and buildings) and climate and environmental forcing variables, which gave us five major categories of data: (i) data on chemical and ecological states of the watercourses, (ii) data characterizing the sampling reaches, (iii) data describing the hydrographic network, (iv) data estimating human activities, and (v) climate and environmental forcing variables.

All of these data have spatial characteristics and they are complex to structure and to inter-connect because of their volume and nature. They are characterized by high heterogeneity due to their origin (values derived from measurements or expert assessments), their value that can be quantitative, semiquantitative or qualitative, and their structure (point, line, surface polygon), as well as because of their temporal variability (sampling duration and frequency). Furthermore, the obtained data are not of the highest quality because of the size, the complexity of the French water network, the changes in parameters and methods. Due to the cost, time and human input required for the sampling campaigns, these imperfect collected data are the only available data. These are general problems for environmental data integrating spatial and temporal aspects from several sources with a different spatiotemporal accuracy (Suteanul, 2010).

2.2. Towards an innovative data quality process

The aim of the Fresqueau project is to link all collected data to

[1] Part of the data was downloaded from the Eau France web portal (http://www.eaufrance.fr/), while the rest was obtained from regional environmental agencies. These data were supplemented by detailed local measurements, e.g. in Alsace (Grac et al., 2011).

build a global assessment of waterbodies. We therefore proposed to design a spatiotemporal database and warehouse for exploring the data and to support the assessment decision process. Environmental data sets are complex, with spatial and temporal aspects, and various precision levels; they also have intrinsic quality flaws. Our approach thus integrates quality processes from the modeling to the implementation step, based on domain expertise. The integration of data quality processes is currently one of the main priorities to help make appropriate decisions. Instead of evaluating the data quality with metrics and integrating data according to cleaning rules as is usually done in the ETL process, we decided to tackle data quality issues in a new data-quality aware framework.

The spatiotemporal information-system model proposed here integrates several data-quality aspects in a dedicated package. According to expert advice, the spatial, thematic, and temporal accuracy of data are supported:

- Spatial accuracy measures relations that may exist between objects' features in order to maintain the relevant spatial relationships between them.
- Thematic accuracy describes the accuracy of the attribute values encoded in the database. For example, values of physico-chemical measurements do not have the same confidence if a valid quantification or detection threshold is included. The hydrological theme is integrated in our approach in order to make more relevant diagnoses of the ecological water status.
- Temporal accuracy measures the extent to which the collected data remains consistent over time.

Based on these definitions of the main quality elements, let us now consider the decision-making context where the expert needs both a vast quantity of information over a time course as well as accurate data to make the right decisions throughout the analysis process. In particular we consider that all data, whether accurate or not must be weighted according to their quality level for each data quality dimension that is being monitored. The quality metrics are not used here for cleaning purposes before being integrated into the data warehouse. Once the dimensions are weighted, the experts can make multiple and more accurate interpretations of the analysis results depending on the data quality level as discussed hereafter.

3. Related work

Data warehousing and OLAP systems allow multidimensional online analysis of huge volumes of data according to the multidimensional model which defines the concepts of facts and dimensions. The facts of a data warehouse are the values of the indicators to be analyzed (Malinowski and Zimanyi, 2008). An example of a data warehouse is described in Mazón and Trujillo (2008), where facts are the product sales of a company in dollars. The following example is an excerpt from the data warehouse described below. DW facts are measurements of physico-chemical parameters over time (e.g. pH, temperature, dissolved oxygen) obtained at sample sites on several watercourses. In a data warehouse, an analysis is performed using an aggregation operation (e.g. sum or average) on the facts. In the example, a possible analysis is the average of measurements calculated per physico-chemical parameter, sample site and month. The results of this analysis are represented in a cube (see Fig. 1(a)). Each dimension of the cube corresponds to a analysis criterion: type of parameter, sample site and month. The cube cells are called measures which store the average measurements for each tuple ⟨parameter, sample site, month⟩. For instance, in Fig. 1(a), the average measurements for the tuple ⟨pH, Site 1, December⟩ are 6.3. In data warehouses, the analysis criteria are structured in hierarchies called dimensions. Fig. 1(b) shows the three dimensions, for parameters, sample sites and time. A data warehouse can produce many analyses by combining different dimension levels. For example, other cubes could be calculated:

- average measurements per watercourse, year,
- average measurements per parameter category, watercourse rank, semester, etc.

Data warehouses generally support *n*-dimensional cubes. Data can be combined to provide previously unknown causal links. Data cubes are explored with OLAP operators that allow navigation within dimension hierarchies and aggregate data, and selection of a subset of warehoused data.

Introducing spatial data into DWs leads to a spatio-multidimensional model where measures can be geometric values and dimension levels can have geometric attributes, thus allowing cartographic visualization of SOLAP queries. A typical spatial relational OLAP (Spatial ROLAP) architecture has three tiers (Fig. 2): spatial DW tier, SOLAP server tier and SOLAP client tier (Bimonte and Miquel, 2010). The spatial DW tier integrates (spatial) data from multiple data sources and manages them using a spatial relational DBMS (e.g. Oracle Spatial) granting scalability and good performance. Before being loaded in the SDW, warehoused data are transformed and cleaned using ETL tools. The SOLAP server implements SOLAP operators that compute and handle spatial data cubes. Finally, the SOLAP client tier provides decision-makers with interactive visual displays (crosstabs, histograms and maps) that trigger SOLAP operators and allow visualization of query results (Bimonte and Miquel, 2010).

Spatial OLAP has been successfully applied in several application domains such as marketing, health monitoring, and agriculture (Nilakanta et al., 2008). Some recent studies have investigated using (S)OLAP for hydrological pollutants analysis. Alexandru et al. (2010) present a multidimensional model for the analysis of natural pollution risks where the pollution value is

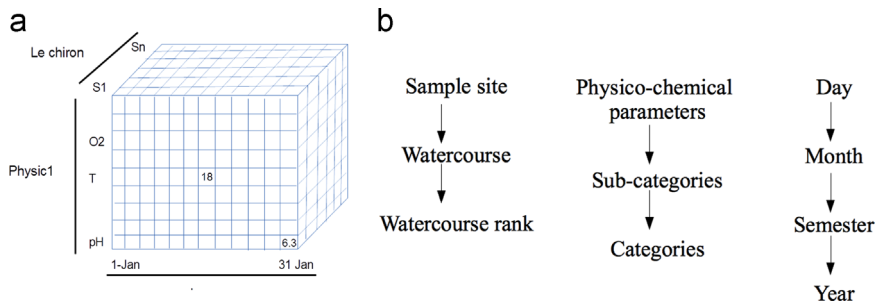

Fig. 1. A simple example of a data warehouse on physicochemical parameters measured in watercourses. (a) Cube on physicochemical parameters. (b) Analysis dimensions.

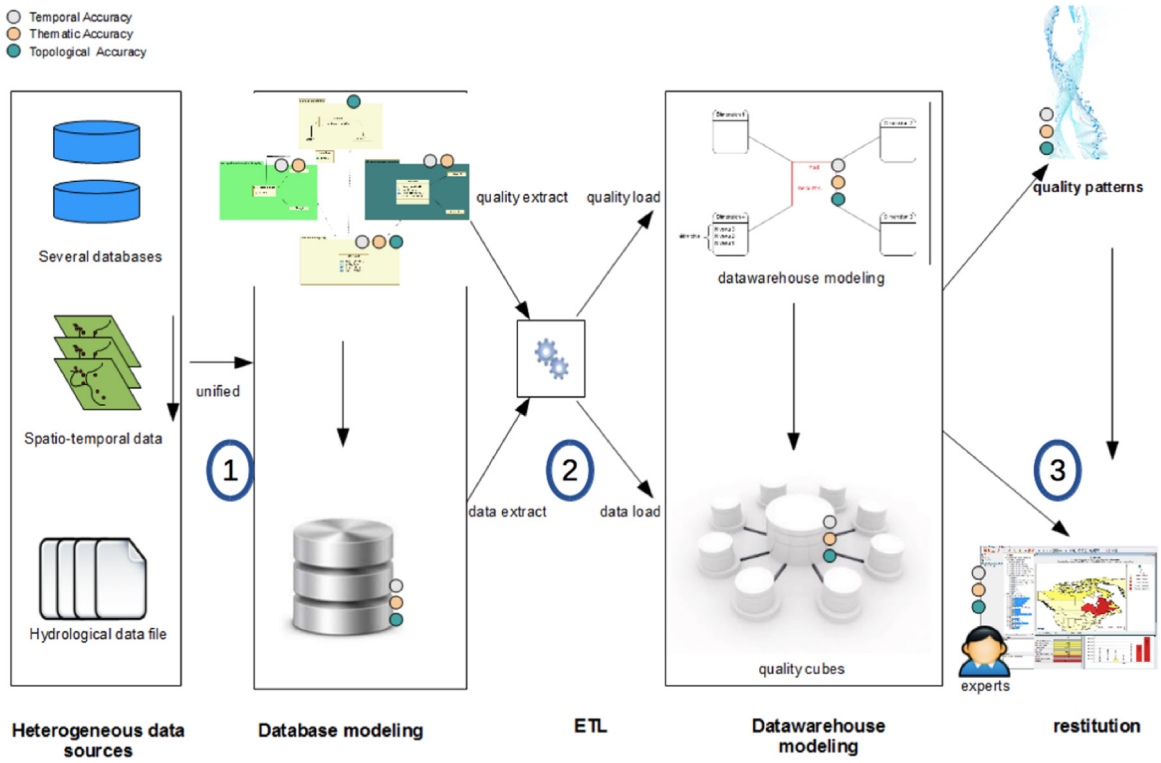

Fig. 2. Data processing within QUIDA global architecture (steps 1–3).

described per pollutant and group of pollutants, in the same way as Vernier et al. (2013) define a SOLAP system for the analysis of agricultural pollutants.

However, in spatial and classical data warehousing systems, data quality has mainly been studied in the context of data cleaning techniques (Bouzeghoub and Kedad, 2002; Jarke et al., 1999). The main objective is to detect and correct errors using declarative ETL operators (e.g., for detection of illegal formats and value replacement). A wide range of techniques are available for detecting and correcting specific value errors, which are often dedicated to specific domains (Rahm and Do, 2000; Lee et al., 1999). In the last decade, numerous cleaning tools have been proposed in the literature and often commercialized. A comparative overview of tools and underlying techniques can be found in Muller and Freytag (2003). However, cleaning techniques cannot be used to define data quality dimensions specific to data warehousing systems or to report them to users.

Only a few studies involve multidimensional modeling of data warehouse quality and focus on reporting real data quality to users. In Zhu and Buchmann (2002), the quality of external data sources is evaluated in order to select the most appropriate ones. They studied and compared different methods used for building quality indicators that are capable of aggregating several quality measures; the syntactic correctness ratio is one of the proposed quality metrics. In Amaral and Campos (2004), the authors build multidimensional cubes to show aggregations of data warehouse quality. In Berti-Equille et al. (2011) and Peralta et al. (2009), a multidimensional data model is used for the analysis of data quality measures. The authors describe two case studies based on customer relationship management (CRM) and health care data warehousing systems where the meta-model is instantiated for data quality analysis and reporting. Syntactic correctness, completeness and record integrity along with other quality metrics have been defined and measured in these contexts based on the goal-question-metric (GQM) paradigm.

4. Quality inside the data warehouse: the QuIDA architecture

Our proposal focuses on a new architecture that fits the quality requirements within an environmental information system. The proposed architecture can be used to transfer data quality indicators from input databases to the spatial data warehouse. The queries are now not only designed from, and submitted to, the different hierarchies, as is usually the case in a multidimensional context, but they can also take the required data quality level into account. The details of the QuIDA architecture are illustrated in Fig. 2, from left to right: (1) several databases and various files are unified within an integrated database including a data quality model; (2) data and quality data are extracted and loaded within quality cubes; (3) data from the warehouse can be explored with data mining methods or can be restituted to the experts with quality information.

Firstly, we propose a modeling step completely tailored for environmental purposes as it pools heterogeneous spatiotemporal data within a unified model (i.e., it can reconcile data from a semantic standpoint). Once the model is unified, data can be properly structured and integrated to improve quality control in the data warehouse. Quality dimensions required by domain experts can be monitored during this step. Secondly, we integrate the data into the spatial data warehouse without using any cleaning rules. ETL processes handle data integration as well as data quality propagation within the spatial data warehouse according to aggregative rules and a weighting function. Finally, we propose a novel way to query cubes using data quality as a query filter. This allows experts to navigate within the different dimension hierarchies but also within the different data quality levels according to their specific accuracy needs.

4.1. Modeling and data integration in the database

In the modeling step, data are compiled within different packages to conserve their original semantics and respect the

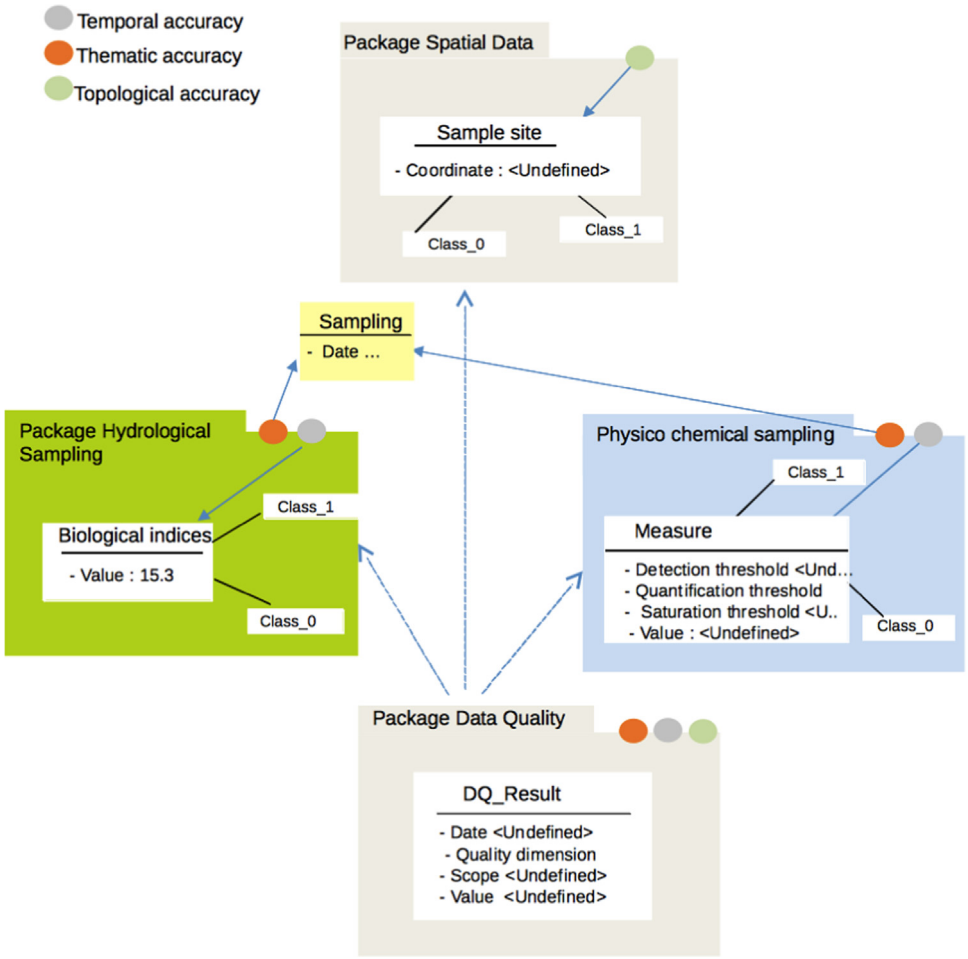

Fig. 3. Simplified data quality model: the three main packages and the quality package.

different viewpoints in the unified model. The main packages are (Fig. 3)

1. The spatial data package links the French water system to the French administrative territorial system which is made up of Regions, Departments, etc. Each watercourse is divided into reaches so that the whole French water system forms a network on which each sample site can be referenced based on its coordinates.
2. The hydrobiological sampling package describes, over a time course, biological indices based on surveys of different species living within the sites.
3. The physicochemical sampling package describes, over a time course, measurements of various physicochemical parameters observed at the sample sites.

The data quality package is based on a spatial data quality literature (Guptill, 2008; Parmar and Goyal, 2012), and, in particular, the ISO 19115 standard (Danko, 2000), that recommends monitoring spatial data quality according to the topological and temporal accuracy. The ISO 19115 standard stipulates that each measure of data quality should be stored in a DQ_Result class, allowing monitoring of data quality changes over time. Data can be traced according to the different data quality elements noted as DQ_Elements classes that represent the data quality dimensions of interest, e.g. completeness, freshness, consistency and accuracy.

Furthermore, each package described in Fig. 3 is tagged with data quality dimensions according to the studied field. We report here on three specific dimensions focused on environmental data quality issues.

- The spatial data package can be used to monitor topological accuracy. We decided to monitor each spatial object using its coordinates to control the topological and positional accuracy.
- The thematic packages (the hydrobiological and physicochemical packages are both tagged with the temporal and thematic labels). Business rules of the studied field are given by experts and then used to define the thematic accuracy.
- Spatial objects and measurements are interdependent and change over time. The temporal quality dimension is monitored systematically in order to analyze their interdependencies correctly,

In a package, each record can be traced using a data quality threshold. Later, the record is given a weight between 0 and 1 according to its corresponding threshold within the process of integration into the quality-aware data warehouse. It is clearly mentioned here that this weight is only used as a quality indicator in order to build a kind of quality, i.e., for each data item on the different quality dimensions.

4.2. Quality cube

The Fact table provides the parameter values according to the different dimension hierarchies. There are three main dimensions surrounding the Fact table (see Fig. 4).

- *Spatial dimension*: the hierarchies are based on the French water

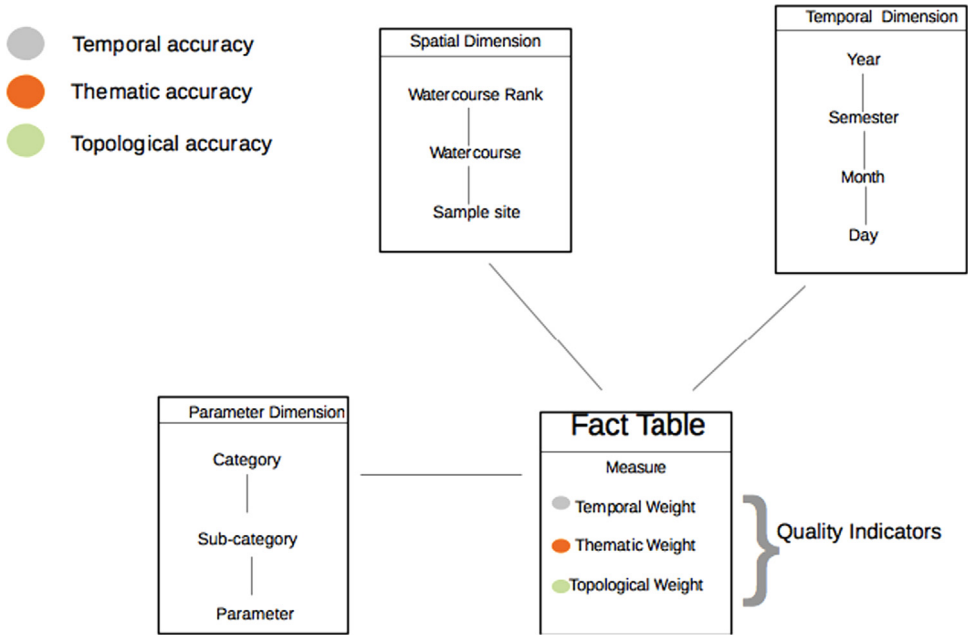

Fig. 4. Spatial quality data warehouse star schema: a measure is a tuple ⟨ parameter, unit, value, date, sample site ⟩. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

system network, a sample site belongs to a watercourse, which in turn belongs to a watercourse rank. Sample sites are weighted according to topological accuracy.

- *Temporal dimension*: Each sample is time stamped with the day, month, and year. The hierarchies on the temporal dimension are constructed according to the different analyses over the time course considered by the experts.
- *Parameter dimension*: This refers to the thematic classification of the different physicochemical parameters in the hydrological field.

Other spatial and thematic dimensions have been developed, e.g. the administrative dimension (a sample station is located in a town or village territory, which is included in a department, etc.). Each record in the Fact table can be queried on various dimensions of the cube to be spatiotemporally analyzed. The value is weighted with the three quality dimensions and the quality weight is considered as a measure directly recorded in the Fact table.

The quality indicators, namely the temporal, topological and thematic quality weights are computed by functions weighting and combining several logical and domain-dependent constraints for consistency and completeness checking. More specifically, temporal accuracy constraints will check if the date and time of the measurement data are non-null values and if they are valid, e.g., not during the week-end or national day or in the middle of the night. As illustrative examples, temporal accuracy of a measurement value on `(date=14JULY2014, time=NULL)` equals 0, whereas the temporal accuracy of `(date=1SEPT2014, time=12am)` is 0.5. The thematic dimension of data quality is evaluated with respect to a set of domain-specific constraints checking the consistency of multiple physicochemical parameters of the measurement. For instance, the percentage of O_2 saturation should satisfy the following equation:

$$\text{SaturationPerc}(O_2) = (O_2 \times 100)/(14.631(0.4112 \times T) + (0.008536 \times T^2) + (0.00009 \times T^3))$$

with T being the water temperature. Various rules have been defined by the domain experts to check the consistency and detect outlying measurement values.

Topological accuracy is computed as the distance between the considered measurement point and the French water network reference, BD Topo.® [2] Topological accuracy is proportional to the distance and corresponds to one of the following levels: high topological accuracy in [.75,1] represented by a green cell for less than 1 m apart from the reference network; medium high for topological accuracy in [.5,.75] represented in yellow when the distance is between 1 and 10 m; medium low for topological accuracy in [.25,.5] represented in orange when the distance is between 10 and 50 m and low topological accuracy in [0,.25] represented in red when the distance is beyond 50 m. This indicates if the measuring point does coincide spatially with a river stretch.

As an example regarding the topological accuracy, the value of a parameter recorded in the Fact table is related to a sample site. If this site has a low topological accuracy level in the water network, it directly impacts the parameter measurement used for the spatial analysis. Indeed, experts attempt to discover how sample sites are interconnected to each other according to the different parameter measurements. The spatial impact that a site can have on another one can only be correctly revealed if its topological accuracy is as precise as possible. Our approach allows experts to consider this viewpoint and analyze data more accurately.

Moreover, experts might be interested in obtaining a better level of topological accuracy in order to gather information that will help them reveal such impacts between different sites. The approach allows them to query data according to a better topological accuracy level when rolling up on the hierarchy, at various aggregation levels. The topological weight is used here as a filter when querying the data cube. Other examples are given in the following section.

5. Results

The global architecture was entirely deployed in an open-source technology using PostgreSQL DBMS which is perfectly suited for spatial data, and the Mondrian server to query the

[2] http://professionnels.ign.fr/bdtopo

Table 1
A spatial case study.

Section	Information
Watershed name	*Saône*
Waterbody name	*Ruisseau la Colombine*
Main stream	*Rhône*
Linear length studied	11 km
Station number	10
Studied parameter	Nitrates

Spatial	Time	Value
−La Colombine	+All time	52,53
−ruisseau la colombine	+All time	52,53
COLOMBINE A CHARCENNE 1	−All time	50,00
	+1995	21,00
	+1996	79,00
COLOMBINE A CHARCENNE 2	−All time	60,50
	+1995	35,00
	+1996	86,00
COLOMBINE A CHARCENNE 5	−All time	52,50
	+1995	30,00
	+1996	75,00
COLOMBINE A CHARCENNE 3	−All time	48,00
	+1995	20,00
	+1996	76,00
COLOMBINE A CHARCENNE 4	−All time	50,00
	+1995	28,00
	+1996	72,00
COLOMBINE A CHOYE 2	−All time	46,00
	+1995	17,00
	+1996	75,00
COLOMBINE A CHOYE 3	−All time	55,50
	+1995	26,00
	−1996	85,00
	−1996-05	85,00
	+1996-05-21	85,00
COLOMBINE A CHOYE 1	−All time	65,50
	+1995	39,00
	+1996	92,00
COLOMBINE A CHOYE 4	−All time	37,00
	+1995	37,00

Fig. 5. Cube without quality weights.

cubes. The current cube focuses on physicochemical data. Its size is about 4 Go.

Experiments were performed on a set of 49 sample sites from the Fresqueau dataset. 712 samples from these 49 sites were integrated within the spatial quality data warehouse in order to foster discussion on our results. Those samples concern 4517 physicochemical measurements obtained at different dates and for different physicochemical parameters. These parameters are divided into two sections according to their role:

- Evaluation of the chemical state based on micropollutants: atrazine, nitrates, diuron, mercury, cadmium, trichlorethylene.
- Evaluation of the physicochemical state based on macropollutants: NH_4^+, NO_2^-, NO_3^-, PO_4^{3-}, total phosphorus, pH, dissolved oxygen (DO), dissolved oxygen percent saturation, biological oxygen demand (BOD_5).

The first example aims to show how data quality weights can help decide which section of a watercourse has to be inspected. Table 1 gives an overview of such data.

The section of the watercourse under study is located in an agricultural environment where there is intensive chemical fertilizers use. We tried to determine the possible impact of nitrate pollution in an upstream section of the waterbody according to the results of the different stations monitored. Each station reveals the state of a section. If nitrates measured at a station reveal an abnormal increase, the monitoring network could decide to inspect the section upstream in order to understand the causes. Since these inspections are very costly, it is important to know the topological accuracy along the water network in order to make the right decision concerning the choice of section to be inspected.

In traditional data warehouses, when asking for the nitrate parameter values according to the spatial dimension, the conventional Fig. 5 is obtained: nitrates show different possible impacts from upstream stations, but it is hard to make any relevant decision other than to inspect the nearest upstream station, with the high risk of having to reconduct another costly inspection. In the spatial quality data warehouse, we can differently analyze the data based on the data quality weight. The previous information is given in Fig. 6 with colored cells used to facilitate the analysis.

The thematic weight shows that nitrate concentrations for each station of the studied waterbody belongs to the valid threshold.

In the following, other examples of quality weight queries and results are given. In Fig. 7, the expert is looking for a spatial trend at different stations based on a parameter measurement ("*What are the stations with a topological weight above* 0.75 *for the Atrazine parameter results*?"). The quality data warehouse can be used to query the spatial data accuracy and strengthen the robustness of the analysis. The analysis remains robust even if, as in this case, the temporal accuracy is not high.

Now (Fig. 8), if the expert is interested in showing a temporal rather than spatial trend in the data and no longer over space, he/she can query in the same way the quality data warehouse by using the temporal weight filter: "*What are the years for which the dissolved oxygen results are related to a temporal weight above* 0.75?". And lastly (Fig. 9), spatiotemporal trend may be highlighted using topological and temporal weights: "*What are the stations for which the topological and temporal weight is above* 0.75 *for the nitrate parameter results*?".

We can also request results according to the different

Spatial	Time	Value	Temporal weight	Topological weight	Thematic weight
−La Colombine	+All time	52,53	0,75	0,47	1,00
−ruisseau la colombine	+All time	52,53	0,75	0,47	1,00
COLOMBINE A CHOYES	+All time				
COLOMBINE A CHARCENNE 1	+All time	50,00	0,75	0,50	1,00
COLOMBINE A CHARCENNE 2	+All time	60,50	0,75	0,50	1,00
COLOMBINE A CHARCENNE 5	+All time	52,50	0,75	0,50	1,00
COLOMBINE A CHARCENNE 3	+All time	48,00	0,75	0,50	1,00
COLOMBINE A CHARCENNE 4	+All time	50,00	0,75	0,25	1,00
COLOMBINE A CHOYE 2	+All time	46,00	0,75	0,50	1,00
COLOMBINE A CHOYE 3	+All time	55,50	0,75	0,50	1,00
COLOMBINE A CHOYE 1	+All time	65,50	0,75	0,50	1,00
COLOMBINE A CHOYE 4	+All time	37,00	0,75	0,50	1,00

Fig. 6. Quality cube. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Spatial	Value	Temporal weight	Topological weight	Thematic weight
SAONE A APREMONT	1,48	0,71	0,75	0,66
SOUFROIDE A MANTOCHE	0,11	0,69	0,75	0,75
MORTE A ANCIER	0,06	0,78	0,75	0,72
CABRI A ANGIREY 2	0,07	0,68	0,75	0,70
MORTE A ST-BROING	0,10	0,75	0,75	1,00
COLOMBINE A CHOYES	0,08	0,68	0,75	0,85

Fig. 7. Filters on the topological weight. Question: "*What are the stations with a topological weight above* 0.75 *for the Atrazine parameter results*?"

Time	Value	Temporal weight	Topological weight	Thematic weight
+1995	6,60	0,75	0,47	1,00
+1997	10,64	0,75	0,67	1,00
+2007	9,95	0,81	0,67	1,00
+2008	10,86	0,91	0,69	1,00
+2009	9,86	0,98	0,70	1,00
+2010	9,91	0,87	0,73	1,00

Fig. 8. Filters on the temporal weight. Question: "*What are the years for which the dissolved oxygen results are related to a temporal weight above* 0.75?"

Spatial	Time	Value	Temporal weight	Topological weight	Thematic weight
SAONE A APREMONT	+1997	10,08	0,75	0,75	1,00
	+1998	10,92	0,75	0,75	1,00
	+2001	10,00	0,75	0,75	1,00
	+2003	8,45	0,75	0,75	1,00
	+2007	11,57	0,75	0,75	1,00
	+2008	11,83	1,00	0,75	1,00
	+2010	13,08	0,96	0,75	1,00
VINGEANNE A RENEVE	+1994	18,13	0,75	0,75	1,00
VINGEANNE A ATTRICOURT	+1994	18,37	0,75	0,75	1,00
SOUFROIDE A MANTOCHE	+2010	36,83	0,94	0,75	1,00
BEZE A BEZE 2	+1998	20,70	0,75	0,75	1,00
BEZE A BEZE 1	+1994	24,03	0,75	0,75	1,00
MORTE A BUCEY-LES-GY 2	+1999	13,60	0,75	0,75	1,00
	+2003	11,95	0,75	0,75	1,00
MORTE A ANCIER	+2007	25,48	0,75	0,75	1,00
	+2008	23,05	1,00	0,75	1,00
	+2009	23,77	1,00	0,75	1,00
	+2010	21,93	1,00	0,75	1,00
MORTE A ST-BROING	+2005	30,25	0,75	0,75	1,00
	+2006	31,00	0,75	0,75	1,00

Fig. 9. Filters on topological and temporal weights. Question: "*What are the stations for which the topological and temporal weight is above* 0.75 *for the nitrate parameter results*?".

Spatial	Time	Value	Temporal weight	Topological weight	Thematic weight
−All	+All time	18,73	0,70	0,66	1,00
+La Saône	+All time	10,46	0,70	0,75	1,00
−La Vingeanne	+All time	17,40	0,66	0,61	1,00
−La Vingeanne d'Oisilly à sa confluence avec la Saône	+All time	16,71	0,67	0,54	1,00
VINGEANNE A TALMAY 3	+All time	16,60	0,68	0,50	1,00
VINGEANNE A CHEUGE	+All time	17,43	0,63	1,00	1,00
VINGEANNE A RENEVE	+All time	18,13	0,75	0,75	1,00
+La Vingeanne de l'Etivau à Oisilly Badin Inclus	+All time	19,92	0,59	0,88	1,00
+La Soufroide	+All time	36,83	0,94	0,75	1,00
+Le Chiron	+All time	21,02	0,45	0,85	1,00
+La Bèze	+All time	24,77	0,61	0,61	1,00
+La Morte	+All time	22,35	0,85	0,66	1,00
+L'Albane	+All time	44,14	0,60	0,50	1,00
+La Rèsie	+All time	30,90	0,56	0,50	1,00
+La Tenise	+All time	26,60	1,00	0,50	1,00
+Pannecul	+All time	38,20	0,38	0,50	1,00
+La Colombine	+All time	52,53	0,75	0,47	1,00

Fig. 10. Accuracy trends in hierarchies. Question: "*What are the quality weights for the nitrate results on the different spatial dimension hierarchies*?".

hierarchies and have a look at data accuracy trend on the spatial dimension of the quality data warehouse "*What are the quality weights for the nitrate results on the different hierarchies of the spatial dimension*?" (Fig. 10).

In the following, we show the advantages of the spatial quality data warehouse using the topological accuracy weight. Fig. 11

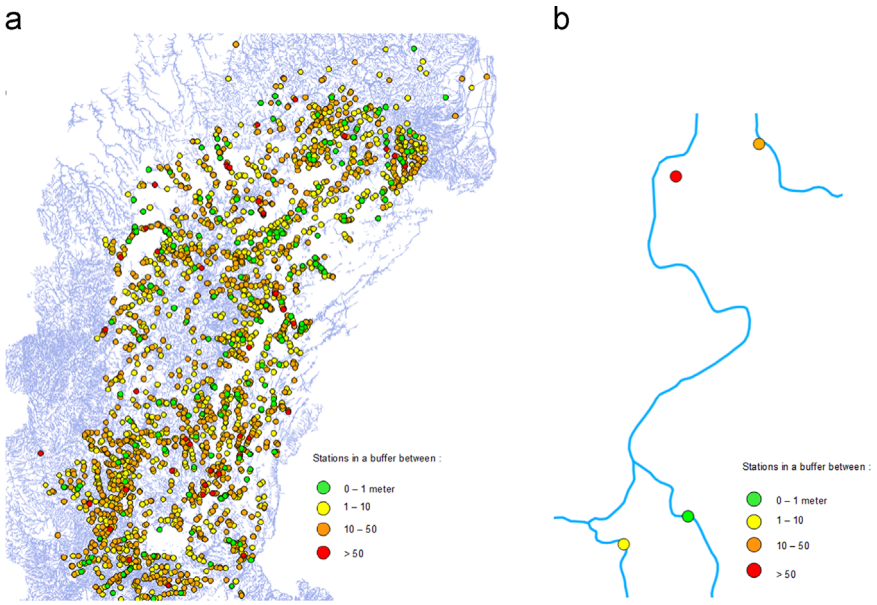

Fig. 11. Topological accuracy restitution. (a) Set of stations. (b) Topological thresholds. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

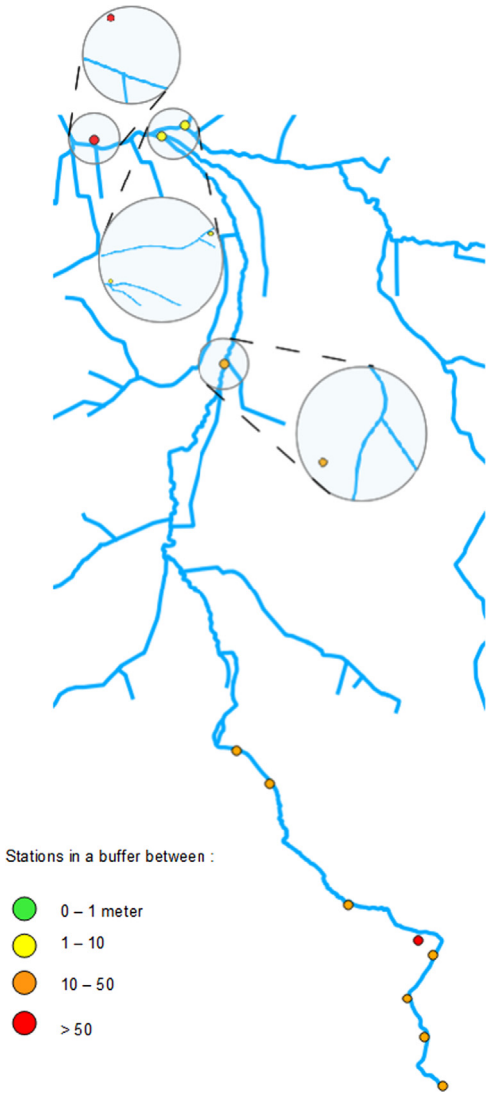

Fig. 12. Topological accuracy for spatial station analysis.

(a) shows the distribution of different stations from which we extracted the set to be monitored. The colors in Fig. 11(a) and (b) represent the spatial restitution of the topological accuracy according to the defined thresholds in the watersheds as *Green* stations in a buffer between the perfect intersection and 1 m; *Yellow* between 1 and 10 m; *Orange* between 10 and 50 m and *Red* beyond 50 m. The specific color code explains the data reliability and helps decisionmakers for quick analysis.

The topological weight highlights different levels of accuracy in the station locations. The topological weight here is really helpful because it clearly shows the difficult choices of section to be inspected since one of the stations is colored in red. Indeed, in Fig. 12, the spatial restitution circles different cases where it is hard to choose which sections of river to inspect. The red case at the top of the figure emphasizes three potential sections of river to be inspected. This allows the operator to decide either to conduct further investigation at that position or to inspect an upstream section with a better topological weight.

6. Discussion and conclusion

Here we have described a novel approach to address data quality issues in the environmental application framework. The originality of this approach resides in the fact that it uses data quality, not for cleaning, but rather for querying data according to their quality level. First, we described the QuIDA architecture to transfer these data quality indicators from the database to the data warehouse. Second, we paid particular attention to the modeling and integration steps according to the data quality dimensions and defined the quality weights as authentic data indicators. Third, we monitored the data quality weights for each record in the data warehouse fact table. And lastly, we implemented our approach on a large dataset concerning French running waters and conducted multiple data warehouse queries using these quality indicators as data cube filters.

Our approach assumes that all data are integrated even though they have a low quality level. As long as it is possible to define and measure the data quality with the indicators, then all the data may be used in a decision-making context. This information is directly entered as a measurement in the fact table for each record and

according to different quality dimensions. However, we could have analyzed data quality within the data warehouse in a different manner. A discrete data quality dimension could have been added, with three or more data quality levels, e.g. imprecise, precise and confident levels.

In our model, the aggregative rules to roll-up or drill-down regarding the quality dimension could not be consistently defined because these hierarchies are not present in our data structures. Moreover, each quality dimension has its own quality level definition, so it was impossible to consider them all within the same dimension. The only way to transfer the data quality and compare different queries on cubes is to directly measure quality weights in the Fact table. Indeed, beyond the data quality reporting discussed previously, experts need to discover relevant rules between the physicochemical parameters and the biological measures in order to better assess the quality of hydroecosystems.

To the best of our knowledge, this is the first study to develop a global solution that considers the data quality level inside the system in the hydroecological field. By proposing this solution, we aim to facilitate the spatiotemporal analysis and help to define efficient and reliable conclusions on data that are often scare and imprecise. The principles of our solution can thus be widely adapted to environmental issues, e.g. in the agriculture domain to spatiotemporally monitor agricultural practices and their measured impacts on ecological processes. Furthermore, we could improve our model and develop a procedure designed to recommend accuracy weights (or part of them) to experts. This step could shorten time-consuming activities for analysts. Finally, starting from the last layer of our architecture, we could propagate and integrate the data quality concept into a further data-mining step, as considered in the Fresqueau project (e.g. Bertaux et al., 2009; Fabrègue et al., 2013).

Acknowledgments

This work was funded by the French National Research Agency (ANR_11 MONU 14 Fresqueau).

References

Alexandru, A., Gorghiu, G., Nicolescu, C.L., Alexandru, C.-A., 2010. Using OLAP systems to manage environmental risks in Dambovita County. Bulletin UASVM Horticulture.

Amaral, G.C.M., Campos, M.L.M., 2004. AQUAWARE: a data quality support environment for data warehousing. In: SBBD, pp. 121–133.

Bédard, Y., Merrett, T., Han, J., 2001. Fundamentals of spatial data warehousing for geographic knowledge discovery. Geogr. Data Min. Knowl. Discov. 2, 53–73.

Bertaux, A., Le Ber, F., Braud, A., Trémolières, M., 2009. Identifying ecological traits: a concrete FCA-based approach. In: Ferré, S., Rudolph, S. (Eds.), 7th International Conference on Formal Concept Analysis, ICFCA 2009, Darmstadt, LNAI 5548. Springer-Verlag, Berlin Heidelberg, pp. 224–236.

Berti-Équille, L., Comyn-Wattiau, I., Cosquer, M., Kedad, Z., Nugier, S., Peralta, V., Si-Saïd Cherfi, S., Thion-Goasdoué, V., 2011. Assessment and analysis of information quality: a multidimensional model and case studies. Int. J. Inf. Q. 2 (4), 300–323.

Bimonte, S., Miquel, M., 2010. When spatial analysis meets OLAP: multidimensional model and operators. Int. J. Data Warehous. Min. 6 (4), 33–60.

Bouzeghoub, M., Kedad, Z., 2002. Quality in data warehousing. In: Information and Database Quality, pp. 163–198.

Danko, D., 2000. ISO 19115 geographic information—metadata, ISO/TC211 geographic information/geomatics. Technical Report, ISO.

Fabrègue, M., Braud, A., Bringay, S., Le Ber, F., Teisseire, M., 2013. OrderSpan: mining closed partially ordered patterns. In: The Twelfth International Symposium on Intelligent Data Analysis (IDA 2013), London, United Kingdom, vol. LNCS 8207, Springer, Berlin Heidelberg, pp. 186–197.

Grac, C., Braud, A., Le Ber, F., Trémolières, M., 2011. Un système d'information pour le suivi et l'évaluation de la qualité des cours d'eau – Application à l'hydro-écorégion de la plaine d'Alsace. RSTI—Ing. Syst. d'Inf. 16, 9–30.

Guptill, S.C., 2008. Fundamentals of spatial data quality. Trans. GIS 12 (1), 161–162.

Jarke, M., Jeusfeld, M.A., Quix, C., Vassiliadis, P., 1999. Architecture and quality in data warehouses: an extended repository approach. Inf. Syst. 24 (3), 229–253.

Kimball, R., 1996. The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, New York 374 p.

Lee, M.L., Lu, H., Ling, T.W., Ko, Y.T., 1999. Cleansing data for mining and warehousing. In: 10th International Conference on Database and Expert Systems Applications, DEXA'99 Florence, Italy, vol. LNCS 1677, pp. 751–760.

Malinowski, E., Zimanyi, E., 2008. Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications. Springer, Berlin Heidelberg, 435 p.

Mazón, J.-N., Trujillo, J., 2008. An MDA approach for the development of data warehouses. Decis. Support Syst. 45 (1), 41–58.

Muller, H., Freytag, J., 2003. Problems, methods, and challenges in comprehensive data cleansing. Technical Report HUB-IB-164, Humboldt University Berlin, Germany.

Nilakanta, S., Scheibe, K., Rai, A., 2008. Dimensional issues in agricultural data warehouse designs. Comput. Electron. Agric. 60 (3), 263–278.

Parmar, V., Goyal, P., 2012. Analysis on axioms of spatial data quality. In: IJCA Proceedings on National Workshop-Cum-Conference on Recent Trends in Mathematics and Computing 2011, vol. RTMC 1.

Peralta, V., Thion-Goasdoué, V., Kedad, Z., Berti-Equille, L., Comyn-Wattiau, I., Nugier, S., Si-Said Cherfi, S., 2009. Multidimensional management and analysis of quality measures for CRM applications in an electricity company. In: ICIQ, pp. 202–215.

Rahm, E., Do, H.H., 2000. Data cleaning: problems and current approaches. IEEE Data Eng. Bull. 23 (4), 3–13.

Suteanul, C., 2010. A scale-space information flux approach to natural irregular patterns: methods and applications. J. Environ. Inf. 16 (2), 57–69.

The European Parliament and the Council, 2000. Framework for Community action in the field of water policy. Directive 2000/60/EC (23 October 2000).

Vernier, F., Miralles, A., Pinet, F., Carluer, N., Gouy, V., Molla, G., Petit, K., 2013. EIS Pesticides: An environmental information system to characterize agricultural activities and calculate agro-environmental indicators at embedded watershed scales. Agric. Syst. 122, 11–21.

Zhu, Y., Buchmann, A., 2002. Evaluating and selecting web sources as external information resources of a data warehouse. In: 3rd International Conference on Web Information Systems Engineering (WISE), pp. 149–160.