Multidimensional modeling and analysis of large and complex watercourse data: an OLAP-based solution

Kamal Boulil, Florence Le Ber, Sandro Bimonte, Corinne Grac, Flavie Cernesson

Ecological Informatics

Multidimensional modeling and analysis of large and complex watercourse data: an OLAP-based solution

Kamal Boulil [a], Florence Le Ber [a], Sandro Bimonte [b], Corinne Grac [c], Flavie Cernesson [d]

[a] Laboratoire ICube, Université de Strasbourg, ENGEES, CNRS, 300 bd Sébastien Brant, F67412 Illkirch, France
[b] Equipe Copain, UR TSCF, Irstea—Centre de Clermont-Ferrand, 24 Avenue des Landais, F63170 Aubière, France
[c] Laboratoire LIVE, Université de Strasbourg/ENGEES, CNRS, rue de l'Argonne, F67000 Strasbourg, France
[d] AgroParisTech—TETIS, 500 rue Jean François Breton, F34090 Montpellier, France

1. Introduction

In the context of the European Water Framework Directive (DCE, 2000), the Fresqueau project funded by the French Agency for research ANR (2011–2014) aims to develop new methods for studying, comparing and exploiting all the available parameters concerning the status of running waters as well as the information describing uses and undertaken measures. More precisely, the project will contribute to the answer to two specific issues: (1) going farther into the understanding of running water functioning through the analysis of taxa that support biological indices, and (2) connecting the sources of pressure and the physicochemical and biological quality of running waters. The first step of the project was the definition of an integrated Database (DB), integrating data from 20 public databases related to water quality assessment. This large DB (2.6 Go) was constructed to provide data analysis and knowledge discovery tools and methods with the necessary data at the appropriate level of detail. This large DB contains data related to five major topics: water quality parameters (bioindices, physicochemical parameters, etc.), hydrographic networks (descriptions of water quality stations and their hydrographic networks), land uses (land cover, flow obstacles, etc.), and contextual factors (climate, etc.).

Current solutions (text files, ad hoc programs, spreadsheet tools such as Open Office and MS Excel, etc.) used by water quality practitioners are not suited to manage and to analyze large volumes of information. Recently, some studies have shown the ease and power of using Data Warehouse (DW) and On-Line Analytical Processing (OLAP) technologies to store and to analyze environmental data (Alexandru et al., 2010; Boulil et al., 2013b; McGuire et al., 2008; Pinet and Schneider, 2010). DWs are databases dedicated to the integration and storage of large volumes of data to support the decision processes of organizations (Inmon, 2005). DWs store decisional data at the finest granularity level and organize the data in a way that facilitates the analysis/aggregation. OLAP tools allow an interactive exploration of DW data at different levels of detail, following a multidimensional approach. These tools build multidimensional data structures having different granularities, called data cubes, by aggregating DW data and provide users with operators for rapid exploration of these data cubes. Data cubes represent the measures/metrics (e.g., temperature) of the subjects analyzed or of the facts in a space with multiple dimensions (analysis criteria for facts such as time and geographic locations, etc.) according to the multidimensional abstraction model. The dimensions are organized into hierarchies

E-mail addresses: kamal.boulil@gmail.com (K. Boulil), florence.leber@engees.unistra.fr (F. Le Ber), sandro.bimonte@irstea.fr (S. Bimonte), corinne.grac@engees.unistra.fr (C. Grac), flavie.cernesson@teledetection.fr (F. Cernesson).

of aggregation levels to allow viewing analysis indicators (aggregations of measures) at different granularities.

DW and OLAP data cubes are generally implemented in relational platforms consisting of four tiers (Boulil et al., 2013b; Malinowski and Zimányi, 2008): the ETL that integrates data from data sources into the DW, the DW (a relational database that stores the finest data), the OLAP server that calculates the data cubes from the DW data and the OLAP client that displays the data cube information using tables and different types of statistical diagrams (pie charts, histograms, etc.) and reports in different export formats. Unlike the OLTP tools, OLAP tools provide end users with easy to use and powerful analysis methods that enable dynamic changes of the analysis perspective and the granularity of data. The OLAP client interface is used to trigger OLAP operations such as Roll-up and Drill-down which, respectively, decrease and increase the granularity of indicator values, and Slice, which returns a sub-cube by applying a filtering condition on one dimension, etc.

In addition to data structures (facts, measures, dimensions, etc.), the definition of analysis indicators is a fundamental part of data cubes. The analysis indicators are computed by aggregating measure values along dimension hierarchies. A simple analysis indicator is an application of a common aggregate function to a measure along all dimensions of the data cube (e.g., the average temperature (Fig. 1)). Common aggregate functions are supported by OLAP tools and DBMSs (e.g., Sum and Count). In contrast, a complex analysis indicator may involve different aggregate functions on different measures and along different dimensions, or simply a complex aggregate function on a measure. Complex aggregate functions (e.g., percentile functions) are not supported by OLAP tools.

In the literature, many multidimensional models (languages) (Abelló et al., 2006; Luján-Mora et al., 2006; Malinowski and Zimányi, 2008; Pinet and Schneider, 2010) and development approaches (Glorio and Trujillo, 2008; Hahn et al., 2000; Pardillo and Mazón, 2010; Romero and Abelló, 2009) have been proposed to model data cubes, but none of these models and approaches has been adopted as a standard. All the existing propositions in the area of multidimensional modeling ignore the definition aspect of analysis indicators; the propositions only allow the design of simple indicators. In Boulil et al. (2013a), we proposed a conceptual and implementation framework for data cubes. This framework is based on standard modeling and implementation languages (UML, OCL, SQL, and MDX) and allows for graphical modeling and automated implementations of data cubes. The conceptual framework is formalized as a UML profile, an extension of UML. Unlike related work, our framework particularly allows the definition of complex analysis indicators using complex and/or multiple aggregate functions.

In this paper, we first show the application of the OLAP technology to the field of water quality assessment. The architecture of the OLAP system that we defined consists of two data cubes, a data cube storing data about the physicochemical water quality and another data cube concerning hydrobiological water quality data, tools that allow their periodical feeding with data from operational data sources (an integrated database and some Excel files) and tools for the OLAP analysis by water quality practitioners. This architecture is based only on free software and can easily be extended with other data cubes (e.g., a data cube for the analysis of morphological data on watercourses that is another important element in the water quality assessment) and software components (other data sources, other data analysis tools such as data mining, etc.). Using some examples, we demonstrate how the OLAP technology, unlike OLTP tools such as Excel, can help water quality practitioners to increase their productivity by offering a series of intuitive interfaces that facilitate and accelerate the multidimensional analysis and understanding of water quality data. We render this possible by defining various analysis indicators and enabling simple (thematic, spatial, and temporal) and combined (spatiotemporal) analysis on multiple scales. Using our framework (Boulil et al., 2013a) that we extend here by complex aggregate functions (e.g., a generic function to calculate all percentiles), we show how to define complex analysis indicators by using these complex functions and by introducing additional analysis dimensions to allow their calculation and also for information rendering purposes. In Boulil et al. (2013a, 2013b), we have shown how to define complex indicators having multiple but only simple aggregate functions. In addition, we propose two strategies to address the heterogeneity of measurement units (one of the main summarizability semantic problems) by (i) transforming source data at the ETL tier, and (ii) by introducing an additional analysis dimension at the OLAP server tier. Semantic and structural summarizability conditions grant the accuracy and the correctness of indicator values if we assume a good DW data quality (please see Boulil et al. (2012) for more details about the quality of OLAP analysis indicators). Finally, this paper constitutes a second application in the environmental domain of our framework for data cubes (Boulil et al., 2013a); a first experiment of this framework in agriculture is presented in (Boulil et al., 2013b). Our framework considerably reduces the development times and effort by automating most of the implementation tasks. Our framework is used in this project to design and implement the water quality data cubes and their analysis indicators.

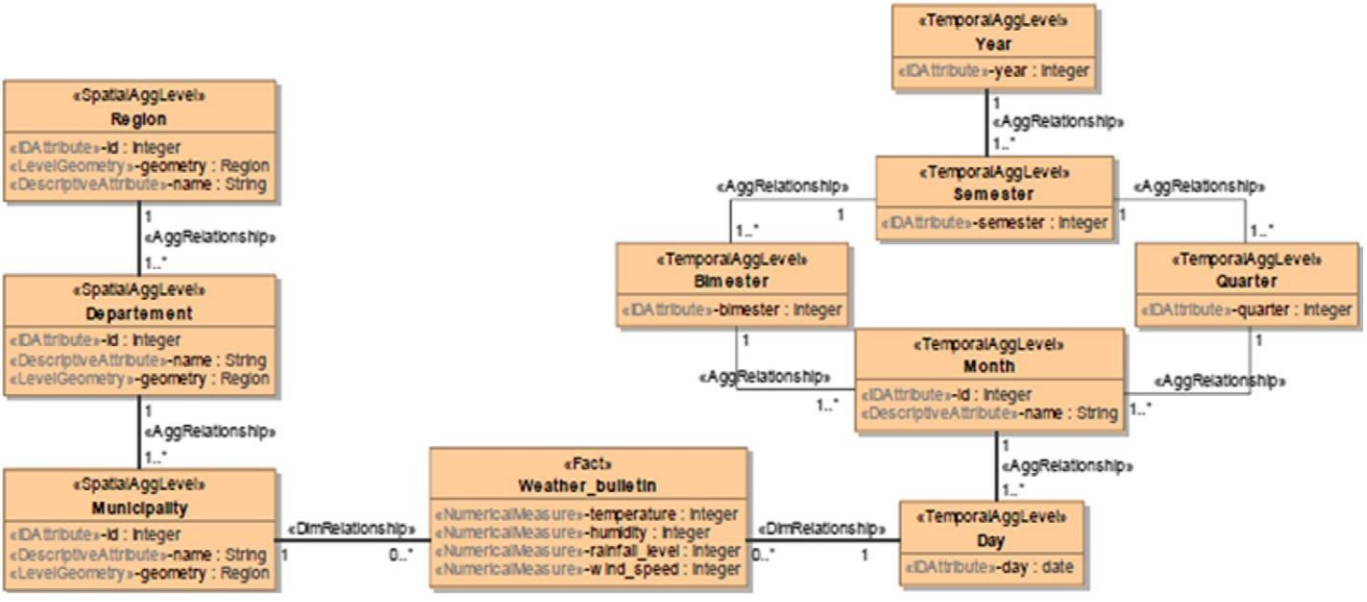

Fig. 1. A multidimensional model for analysis of weather.

The remainder of this paper is organized as follows. Section 2 introduces the main DW and OLAP data cube concepts and also presents the main concepts used in the Fresqueau project of our standard-based framework for the definition of data cubes. Section 3 describes related work and particularly discusses existing data cubes in the field of water quality assessment. Section 4 presents the Fresqueau project. In Section 5, we show how our framework is used to define two data cubes for water quality data analysis and particularly, how to define complex indicators and to address heterogeneity of measurement units. Section 6 presents examples of thematic, spatial, temporal and spatiotemporal OLAP analysis on the data cubes that have been built to demonstrate how OLAP tools can help water quality specialists to understand the water quality data rapidly. Finally, Section 7 concludes the paper with future work.

2. OLAP and DW: main concepts

In addition to data mining tools, Data Warehouses (DWs) and On-Line Analytical Processing (OLAP) tools are main Business Intelligence (BI) technologies (Kimball and Ross, 2002).

A DW is defined as "a collection of subject-oriented, integrated, non-volatile, and time-varying data to support the decision-making processes of an organization (Inmon, 2005)". A DW often integrates and stores large datasets from multiple and heterogeneous data sources, and the DW information is generally accessed in a read only way and can have different values associated with different time instants or periods. DW also organizes data according to the main analysis subjects of the organization in way that facilitates data analysis tasks. For example, a DW of a retail company should contain data about analysis subjects such as sales, purchases, inventory management, etc. DWs are implemented mostly in terms of relational databases following the well-known star schema (Kimball and Ross, 2002). This schema defines two main data structures: fact tables that store values of measures or metrics of the analysis subjects and dimension tables that store data related to analysis criteria.

OLAP tools are a type of software allowing end users to explore DW data interactively, multidimensionally and at different levels of detail (Codd et al., 1993; Kimball and Ross, 2002). On the basis of some metadata (the OLAP schema), these tools transform DW data into strategic information: different data analysis indicators that can be viewed at different granularities and from different analysis perspectives. Indicators, granularities and/or analysis perspectives can be changed rapidly using OLAP operators that handle DW data in terms of data cubes. A data cube is a multidimensional representation of data where cells represent measure values, and coordinates represent analysis criteria values. In a flattened representation, a data cube corresponds to a table with multiple entries where axes (columns and rows) represent analysis criteria, and cells represent measure values. Data cubes can be implemented in three main ways: (a) pre-computation and storage in optimized multidimensional arrays (Multidimensional OLAP (MOLAP)); (b) storage in relational databases (ROLAP); and (c) Hybrid OLAP (combination of ROLAP and MOLAP). The relational implementation presented in Section 2.3 remains the most dominant implementation.

Decision makers explore these data cubes by navigating through dimension hierarchies and performing OLAP operators. Common OLAP operators are the following: "Slice", which defines a selection on one dimension of the cube; "Dice", which performs a selection on two dimensions or more; "Roll-up", which decreases the granularity of the measure values by aggregating them along a dimension hierarchy; and "Drill-down", which is the reverse of Roll-up.

In Table 1, we present a brief comparison between the OLAP and OLTP (On-Line Transaction Processing) technologies by highlighting their main differences. Unlike OLAP that operates at the tactical and strategic levels, OLTP is a class of transaction-oriented information systems that support operational level daily tasks.

In the rest of this section, we will describe the abstraction model on which data cubes are based (Section 2.1), the UML-based framework we proposed in Boulil et al. (2013a) to design and to implement data cubes (Section 2.2) and the components of the common adopted architecture for their implementation (Section 2.3).

2.1. Multidimensional model

DWs and data cubes are based on the multidimensional model (Abelló et al., 2006; Malinowski and Zimányi, 2008). This model allows for representing decisional data (strategic information) according to the point of view of the decision makers by organizing it multidimensionally into facts and dimensions. Facts represent subjects of analysis and are described by attributes (generally numerical) called measures. For example, in the model represented in Fig. 1, the facts "weather bulletins" are described by temperature, rainfall level, humidity and wind speed measures. Dimensions represent measure analysis criteria and allow for viewing measures from different perspectives. In the example of Fig. 1, weather measures are analyzed according to time and location dimensions. A dimension may consist of multiple hierarchies. For example, the time dimension has two hierarchies, one grouping months by bimesters and another grouping the months by quarters. Each hierarchy defines a way of (or a support for) the measure aggregation by organizing dimension data into different granularities called aggregation levels, e.g., the former time hierarchy consists of levels Day, Month, Bimester and Year. Dimension hierarchies allow for viewing indicators (measure aggregates) at different granularities. For example, the average temperatures, the maximum temperatures, etc., can be viewed by day, month, etc. Indicators are computed by applying aggregate functions (e.g., Sum, Average, Minimum, and Maximum) to measure values. For example, the average temperature is computed using the average function (Avg) on temperature values, the maximum temperature is computed using Max, the Minimum humidity is calculated by applying Min on humidity values, etc.

One of the outstanding properties of the multidimensional is that it allows a simple user to view all possible OLAP queries. Each query corresponds to a combination of aggregation levels (at most, one from each dimension) and one analysis indicator. With n dimensions D1,…, Dn each having Mi aggregation levels and P indicators, we have $M1 \times M2 \times \ldots \times Mn \times P$ combinations (the number of possibilities is

Table 1
Main differences between OLTP and OLAP. Based on (Malinowski and Zimányi, 2008).

	OLTP	DW-OLAP
Purpose/usage	Support of operational tasks	Support of decisional tasks
Users	Numerous (thousands), operatives	Less numerous (hundreds), analysts and decision makers
Usage pattern	regular, predictable, and frequent (every day)	Irregular, not predictable and less frequent
Data	Very detailed (secondary data), current (an average time horizon of 60 to 90 days) and small amounts of data	Less detailed (monthly data), historical (an average time horizon of 5 to 10 years) and large amounts of data
Data model	Normalized (does not allow data redundancies) and optimized for transaction processing performance requirements	Denormalized and optimized for online analytical processing performance requirements
Operations on data	All operations (read, add, modify and delete)	Generally only read and add
Queries	Transactional: access to a small number of records (hundreds)	Analytical: access to and aggregation of a large number of records (millions)

infinite if we consider all filtering conditions that can be applied to aggregation level attributes). For example, for the multidimensional model of Fig. 1, with only one aggregate function (one type of indicator) we have 72 ($3 \times 4 \times 6 \times 1$) possible queries such as "what is the average temperature per month and municipality?", "what is the average temperature per semester and municipality?", etc.

2.2. An overview of our standard-based framework for spatial data cubes

Object-Oriented (OO) models are used heavily for data modeling (Abelló et al., 2006) because they are very expressive and represent static and dynamic aspects of complex applications better. The Unified Modeling Language (UML) is the standard language for OO modeling (OMG, 2011). Because using the UML formalism is time consuming and not sufficient to design DW applications, many authors have proposed UML-based multidimensional models (extensions of UML to represent DW concepts at the conceptual design stage (Boulil et al., 2013a), but to date, no standard model has emerged. Many of these multidimensional models are defined as UML profiles that are a UML extension mechanism that allows adapting the UML metamodel to specific platforms or domains (e.g., conceptual design of environmental DWs). A UML profile consists of a set of stereotypes, tagged values and constraints. Stereotypes are specializations of UML metaclasses (e.g., a specialization of the "class" metaclass) that are rendered by an icon or a name enclosed by bbNN. Tagged values represent the properties of stereotypes and are rendered as tagged value name = 'value'. Constraints are used to formalize the stereotypes and tagged values by capturing all their domain semantics, then preventing the arbitrary use of the profile by designers.

In Boulil et al. (2013a), we proposed a UML profile for conceptual design of spatial data cubes. This profile allows for representing all classical concepts of data cubes such as the facts (using the bbFactNN stereotype), measures, dimensions, hierarchies, aggregation levels, aggregation relationships, relationships between facts and dimensions, etc. The profile defines useful specializations of information allowing for better quality control and automated implementations. For example, dimensions and aggregation levels are classified into three types: thematic, tempo- ral and spatial, measures into numeric, spatial, etc.

One of the outstanding advantages of our profile is that this profile allows the representation of simple and complex aggregations/indicators. First, we distinguish between the measure and analysis indicator concepts. A measure is defined as a fact attribute that can be subjected to different aggregations (Sum, Min, etc.). An indicator (bbIndicatorNN stereotype) is viewed as a result of a measure aggregation. For example, the average population in France is calculated by applying the average aggregate function to "population" measure. In this way, we can associate different analysis indicators with the same measure. To allow the definition of complex indicators, we formalize the concepts of the aggregation rule and the aggregate function as UML operations. An aggregation rule is defined as an application of an aggregate function (bbaggregatorNN) to a measure among all dimensions of a cube (bbAggRuleNN), some dimensions of the cube ((bbDimensionAggRuleNN), some hierarchies, or between two aggregation levels. For the "Average population" indicator, we have one simple aggregation rule (bbAggRuleNN) that is "Average among all dimensions". A basic indicator (bbBasicIndicatorNN) is defined as a set of aggregations that apply to one measure. For example, the "Average population" is a basic indicator. A derived indicator is defined as an expression over basic indicators and may concern many measures.

In Table 2, we summarize all necessary stereotypes to understand the cube models of Section 5.1.

Finally, our profile for DWs has been implemented with a UML-based tool called MagicDraw.[1] This implementation allows designers to design the DW conceptual model graphically using our UML profile and to check its validity (the absence of errors and contradictions in the model). We have also developed a code generator to transform DW models automatically and designed our profile into implementations (DW and OLAP physical schemas).

Table 2
Main stereotypes of our UML profile for DWs.

Stereotype	Specialization of	Semantics
bb Fact NN	Class	A fact class
bb NumericalMeasure NN	Property	A measure having a numeric type
bb SpatialDimension NN	Package	A dimension that contains spatial information (locations of facts)
bb TemporalDimension NN	Package	A dimension that contains temporal information (time instants or time periods of facts)
bb ThematicDimension NN	Package	Dimension that contains only thematic (nonspatial and nontemporal) information
bb SpatialAggLevel NN, bb TemporalAggLevel NN, bb ThematicAggLevel NN	Class	The different levels of dimensions that contain spatial, temporal or thematic information.
bb IDAttribute NN	Property	An attribute used to identify aggregation level instances (members)
bb DescriptiveAttribute NN	Property	An attribute used for member rendering in the application user interface such as a parameter name.
bbBasicIndictorNN	Class	Analysis indicator related to one measure (e.g., average population)
bbAggregatorNN	Operation	An aggregate function (e.g., Avg, Sum, etc.)
bbAggRleNN	Operation	An application of an aggregate function on a measure along all dimensions of a cube

2.3. ROLAP architecture

The classical relational OLAP architecture is composed of four tiers (Boulil et al., 2013b; Malinowski and Zimányi, 2008): ETL, data storage tier, OLAP server and OLAP client.

The ETL (Extract, Transform and Load tools) tier generally consists of programs that extract data from operational internal and external data sources, which integrate them (unify their schema) and periodically load them into the DW.

The data storage tier contains an organization DW and/or several data marts, and some metadata generally used for system administration tasks. The organization DW contains all data at the finest granularity level needed for all analysis needs/subjects of the organization. The organization DW can also be viewed as a set of linked data marts. Data marts (DMs) are small DWs that can contain data related to a sub-set of analysis subjects or a group of end users. Data marts can be fed with data sources or DW data.

DWs and DMs are managed using a Relational DataBase Management System (DBMS) such as PostgreSQL. The DW/DM data can be structured following three schema types: star, snowflake or starflake. In the star schema, every dimension is represented by one table containing all its aggregation level attributes as columns. In the snowflake schema, dimensions are normalized, and each aggregation level is mapped into one table. The starflake schema combines the two representations by normalizing some dimensions or parts of dimensions and denormalizing others. In all of these schema types, each analysis subject (fact) is represented by one table that references the dimension tables by using foreign keys. The well-known constellation schema (a constellation of stars, or snowflakes, or starflakes) is obtained when the MD model is composed of two or more cubes, which eventually share some dimensions.

The choice between normalizing and denormalizing dimensions is often based on the storage cost and the expected query performance.

[1] http://www.nomagic.com.

In contrast to denormalized dimensions, normalized dimensions are easy to maintain and optimize the storage space; however, normalized dimensions decrease the query performance because many joins need to be performed when executing queries.

The relational OLAP (ROLAP) server (e.g., Mondrian) builds data cubes from DM or DW data and implements OLAP operators to handle and to navigate rapidly through these data cubes. Usually, the data cubes (SOLAP Server models) are defined by means of a graphical wizard or using XML files.

Finally, OLAP clients (e.g., JRuBiK) provide users with user friendly and interactive interfaces that trigger OLAP operators and allow the visualization of OLAP query results in the form of pivot tables, different statistical diagrams, tree-maps, etc.

3. Related work

DW and OLAP technologies were developed successively in the 1990s to support decision-making processes in organizations better by allowing integration and storage, multidimensional and multi-scale analysis of large data volumes (Kimball and Ross, 2002). These BI technologies were applied in many domains: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). However, these technologies do not allow for spatial analysis.

New BI technologies, namely, Spatial Data Warehouses (SDWs) and Spatial OLAP (SOLAP), have been introduced to take advantage of the spatial analysis potential of increasing geo-referenced data volumes generated by different technologies (e.g., sensor networks, remote sensing systems). These spatial BI technologies extend DW and OLAP with new data structures (e.g., spatial dimension), aggregate functions (e.g., spatial union) and operators (e.g., spatial drill-down). SDWs are defined as collections of spatial and non-spatial data that support spatial multidimensional analysis (Stefanovic et al., 2000). SOLAP is a class of software tools that allow spatiomultidimensional analysis of SDW data; they combine OLAP and GIS functionalities to provide end users with cartographic, multi-dimensional and multi-scale visualizations of the information (Bédard et al., 2007).

A number of studies apply (S)OLAP to decision support in domains such as marketing, public health monitoring, transportation planning, agriculture, environmental risk management, etc. (Bédard et al., 2007; Miquel et al., 2010). However, only Chen et al. (2007), McGuire and Gangopadhyay (2006), and Wang and Guo (2013) investigate this technology in the specific domain of water quality management. McGuire and Gangopadhyay (2006) present a multidimensional data model that allows the analysis of only one hydrobiological water quality parameter (The Fish Index of Biotic Integrity) at multiple spatial resolutions. The data model was implemented in a relational database management system and linked with a GIS that provides users with visualizations of data on different spatial scales. As the authors exploit a GIS (ArcGIS), using this solution may require a certain level of experience in geographic systems. Additionally, using the proposed solution in other application domains can be difficult because no detail is provided for the general architecture. McGuire et al. (2008) propose an SDW design methodology that is based on the four-step methodology of Kimball and Ross (2002). This methodology is applied in the design of an SDW for ecological research (a research question related to freshwater ecology and analysis of biological sampling results). Chen et al. (2007) concentrate on the integration and propose an integration system/architecture of water quality government repositories that, unlike existing works, supports both deep and shallow integration approaches and exploits semantic relationships among data sources using semantic networks (which assist users in locating related sources for their informational needs). Wang and Guo (2013) present a water quality 2.0 OLAP system designed for the South Water Resources Bureau (TSWRB) of Taiwan: the technology Web 2.0 was adopted to integrate qualified data resources, and an OLAP was designed to analyze water quality data from distributed resources.

In the domain of environmental risk management, some recent papers investigate using (S)OLAP for hydrological pollutants analysis. The authors Alexandru et al. (2010) study the analysis of natural pollution risks presenting a multidimensional model where the pollution value is 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. Pollution has been addressed also by Radulescu and Radulescu (2008), by defining classical pollutant value measures and some risk alert measures: the number of values that exceed the alarm level for a pollutant or a category of pollutants and the number of values that exceed the maximum admissible concentration level for a pollutant or a category of pollutants. In Boulil et al. (2013b), we developed an OLAP system to store and to analyze pesticide transfer data generated by a simulation model called MACRO, to validate the model and compare its different versions. The use of DW and OLAP technologies for the analysis of environmental simulation model results has particularly been investigated in Mahboubi et al. (2013, 2010).

In the domain of public health monitoring, Bédard et al. (2003) propose an application that allows for SOLAP analysis of cancers/diseases, deaths and hospitalizations of individuals following disease/death causes, sex, age, date (time), and location and using meaningful indicators (comparative figures). This work provides a comparison between SOLAP and traditional GIS technologies. Scotch and Parmanto (2006) propose the Spatial OLAP Visualization and Analysis Tool (SOVAT), which combines OLAP and GIS technologies, allowing for spatial and numerical queries. They show an application of this tool in Community Health Assessment research. Datasets concerning cancer, incidence, birth, death, etc., are analyzed following dimensions age, diagnosis, race, sex, etc. Finally, SOLAP applications in transportation planning were presented in Shekhar et al. (2002).

4. Fresqueau project

The European Water Framework Directive (DCE, 2000) was imposed to preserve or restore the good condition of water bodies. The European Water Framework Directive also underlined the need for new tools able to process a large amount of complex information, to assess the functioning of water bodies and the effects of actions that have been undertaken. Actually, the evaluation of water bodies is conducted using biological quality elements, based on macroinvertebrates, oligochaeta, fishes, diatoms or macrophytes. Five French bioassessment indices can be used, based on macroinvertebrates, oligochaeta, fishes, diatoms or macrophytes. For each of the indices, normalized protocols exist for (i) sampling, (ii) identification and counting animals or plants, (iii) calculating intermediate metrics and a final index. For instance, for macroinvertebrates, respectively: (i) XP T 90-333 norm (AFNOR, 2009), (ii) XP T 90-3888 norm (AFNOR, 2010), and (iii) NF T 90-350 norm (AFNOR, 2004) are used to calculate index IBGN.

Physical or chemical anthropogenic degradation is also followed, thanks to numerous parameters (especially macropollutants and micropollutants).

Therefore, on each sampling reach (measurement station), during each year, numerous data on watercourse state are produced: (i) biological data: faunistic and floristic lists, metrics and indices, (ii) at least six series of water analyses for each macropollutant, (iii) analysis of different micropollutants, and (iv) chemical and ecological states according to the level of expertise of these results. Data characterizing sampling reaches or stations, describing the hydrographical network and habitat degradations, complete the previous data.

Furthermore, data estimating human activities (land use and wastewater treatment plants), climate and environmental forcing variables have also to be considered. Finally, there are five major categories of

data: (i) data on chemical and ecological states of the water courses, (ii) data characterizing sampling reaches or measurement stations, (iii) data describing the hydrographical network, (iv) data estimating human activities, and (v) climate and environmental forcing variables.

The Fresqueau project funded by the French Agency for research ANR (2011–2014) aims to develop new methods for studying, comparing and exploiting all the available parameters concerning the status of running waters as well as the information describing uses and undertaken measures. More precisely, the project will contribute to the answer for two specific issues: (1) going farther into the understanding of running water functioning through the analysis of taxa that support biological indices, and (2) connecting the sources of pressures and the physicochemical and biological quality of running waters.

To achieve these objectives, an Information System (IS) has been designed that provides knowledge discovery and data analysis tools such as OLAP and different data mining algorithms with necessary data. The global architecture of the IS is shown in Fig. 2. The integrated Fresqueau database integrates datasets collected from 20 public databases having different access protocols, different use rights and different formats (Lalande et al., 2013). The integration issues are beyond the scope of this paper. These public databases are owned and provided by 12 public institutions such as water agencies, the National Geographic Institute (IGN), the ministries of agriculture and ecology, the European Environmental Agency (EEA), and research units, etc. For example, physicochemical and biological data are produced by French Water Agencies in North-East (Rhin-Meuse watershed) and South-East (Rhône-Méditerranée watershed) of France.

The Fresqueau integrated DB model consists of 7 packages (Fig. 3). We have Physical Chemistry (water quality physicochemical parameters and their values), Hydrobiology (bioindices, taxon lists, etc.), Hydromorphology (physical characteristics of watercourses such as dimensions and shapes of river beds, substrate characteristics, state banks, etc.), Land Uses (land cover, flow obstacles, wastewater treatment plants (WWTPs), etc.), Flows and Climate Packages, which are all connected to the Hydrographic network package that contains descriptions of water quality stations and descriptions of their hydrographic networks (a station (or sampling reach) is a point located along a watercourse-segment that is a line located within a watercourse; the watercourses are related to watersheds of different levels represented by polygons, etc.). Finally, the integrated DB is implemented using PostgreSQL DBMS, and its total size is approximately 2.6 Go.

In what follows, we focus on the definition of data cubes for the OLAP analysis of physicochemical and hydrobiological Fresqueau DB parts.

5. Definition of data cubes for water quality assessment

In this section, we introduce two data cubes for water quality assessment: a data cube for physicochemical data and another data cube for hydrobiological data. In Section 5.1, we present their conceptual design using our UML profile for data cubes. In Section 5.2, we describe their implementation in a ROLAP architecture. In Section 5.3, we present practical solutions to address measurement unit heterogeneity and to define complex indicators; our solutions are discussed and compared to existing solutions in related work.

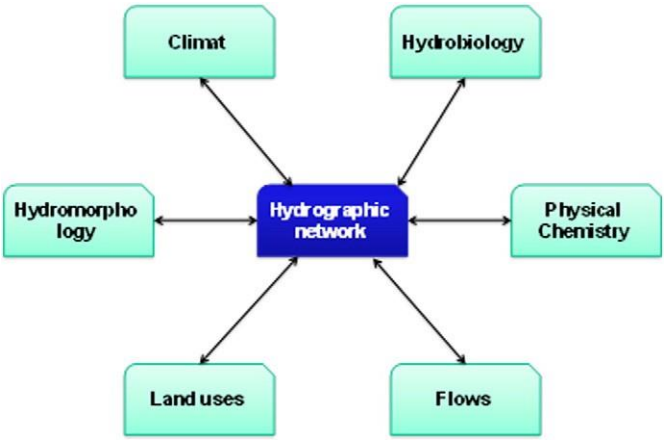

Fig. 3. The Fresqueau integrated DB—main pancakes.

These data cubes have been defined with the help of domain experts (mainly hydrobiologists) who are the end users of the OLAP system. The hydrobiologists have mainly chosen the hierarchies, the interesting dimension levels at which they want to visualize data, and the way data had to be aggregated to compute the desired indicators.

5.1. Data cube conceptual design

In the literature, many multidimensional languages have been proposed to represent data cubes at the conceptual level. In this Section, we present the conceptual models of the two data cubes specified using our UML profile (Sections 5.1.2 and 5.1.3), after briefly introducing the modeling approach we followed to identify the dimensions, measures and analysis indicators of the data cubes (Section 5.1.1).

5.1.1. Modeling approach

In the literature, many DW development approaches have been proposed (Romero and Abelló, 2009). These approaches fall into three main categories:

(1) Data-driven, where the DW or the data mart schema is derived from the schemas of the data sources. This category of approaches guarantees capturing all the analysis potential of the data sources and populating the resulting DW, but the user analysis needs are not considered or are only partly considered.
(2) Requirement-driven approaches start with determining the user analysis requirements to map them onto data sources later. The risk with this category of approaches is that populating of all DW parts is not guaranteed.

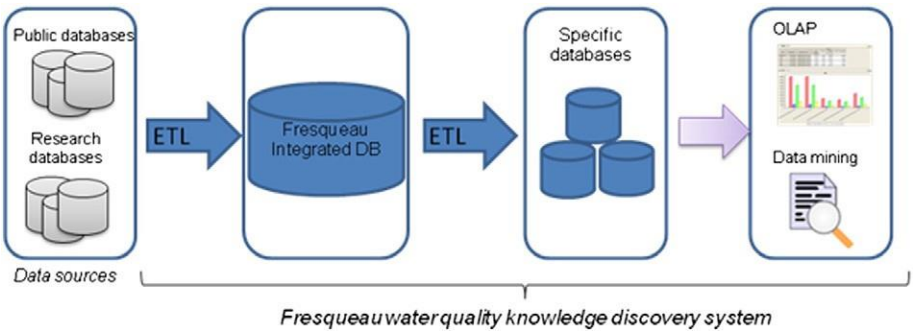

Fig. 2. The Fresqueau IS global architecture.

(3) Hybrid-approaches propose to combine both paradigms to design the DW schema from the data sources but bearing in mind the end user requirements.

To define the data cube models, we follow a hybrid approach. We start by identifying and collecting informational needs of water quality practitioners using classical methods for elicitation of user needs such as interviews, forms, etc. At the same time, we verify the existence of necessary data in the Fresqueau integrated DB to reply to those analysis needs, if not the data availability in other accessible data sources such as Excel files. This verification is performed to avoid defining unusable data cube schemas for which data are not available. From these two elements, user analysis needs and available data, we identify the dimensions, facts, measures and indicators of each data cube.

5.1.2. Physicochemical data cube conceptual model

This cube allows for OLAP analyzing of results of physicochemical samples. The multidimensional diagram of this cube designed using our UML profile is shown in Fig. 4. This model defines one measure ("the value of the physicochemical parameter") expressed in different measurement units (microgram per liter, centimeter per minute, cubic centimeter, gram, gram per meter squared, etc.). This measure is analyzed according to seven dimensions:

(1) Parameter dimension. A thematic (nonspatial and nontemporal) dimension that contains information about water quality physicochemical parameters that are organized in a hierarchy: physicochemical parameters (e.g., glyphosate) are gathered into sub-categories (e.g., pesticides) and sub-categories (e.g., pesticides) into categories (e.g., micropollutants) to allow summing of water quality measures by parameter, sub-category or category.

(2) Station dimension. A spatial dimension that contains data characterizing measurement stations of water quality. Water quality stations (for short, we will use the term station in the rest of this paper) are represented by points in space. This dimension is very important because it allows calculating spatial distributions of water quality measures. The dimension is organized into many spatial hierarchies:

- the administrative hierarchy "Station b Municipality b Department" that groups stations into municipalities and then departments (French administrative divisions).
- the hierarchy "Station b Waterbody b Hydroecoregion_1" that groups stations into water bodies and then into hydroecoregions of level 1. A hydroecoregion of level 1 (e.g., Alsace) is a geographic region with specific climate, geological and hydrological characteristics.
- the hierarchy "Station b Waterbody b Fr_Type" that groups stations into water bodies and then into water body French types that are mainly combinations of two types of information about water bodies and watercourses, their size (small, medium, etc.) and their hydroecoregion of level 2 (e.g., Rhin).
- the hierarchy "Station b Waterbody b Modification_Type" (cf. Fig. 5) that groups stations into water bodies and then into modification categories (natural, artificial, heavily modified, etc.). In response to the second project objective (cf. Section 4), this hierarchy models some sources of human physical pressures on water bodies (dams, channeling, etc.) and enables viewing their influence on water quality indicators.
- the hierarchy "Station b Watercourse b Watercourse rank" that groups stations into watercourses and then into ranks (stream orders that are calculated using a method reverse to the Strahler method (Strahler, 1957)).
- the hierarchy "Station b Watercourse b Watershed_3 b Watershed_2 b Watershed_1" that groups stations into watercourses, then into watersheds (drainage basins) of level 3 (e.g., "Le Sânon": a small waterstream), watersheds of level 2 (e.g., "La Meurthe": a large waterstream), and finally into watersheds of level 1 (e.g., "Le Rhin": a major waterstream).

All of these hierarchies have an "All" aggregation level (containing and aggregating all their dimension members) and are defined to permit different aggregations of water quality measures at various spatial scales and units.

(3) Time dimension. A temporal dimension that contains dates of samples. The temporal dimension defines two hierarchies: "Day b Month b Bimester b Semester b Year" and "Day b Month b Trimester b Semester b Year". This dimension is very important because it allows temporal aggregation and analysis of water quality measures.

(4) Support dimension. A thematic dimension that describes the type of the sampled element (water, sediments, etc.). This information is organized into a hierarchy of two levels: analyzed fractions (e.g., raw water) grouped into supports (e.g., water).

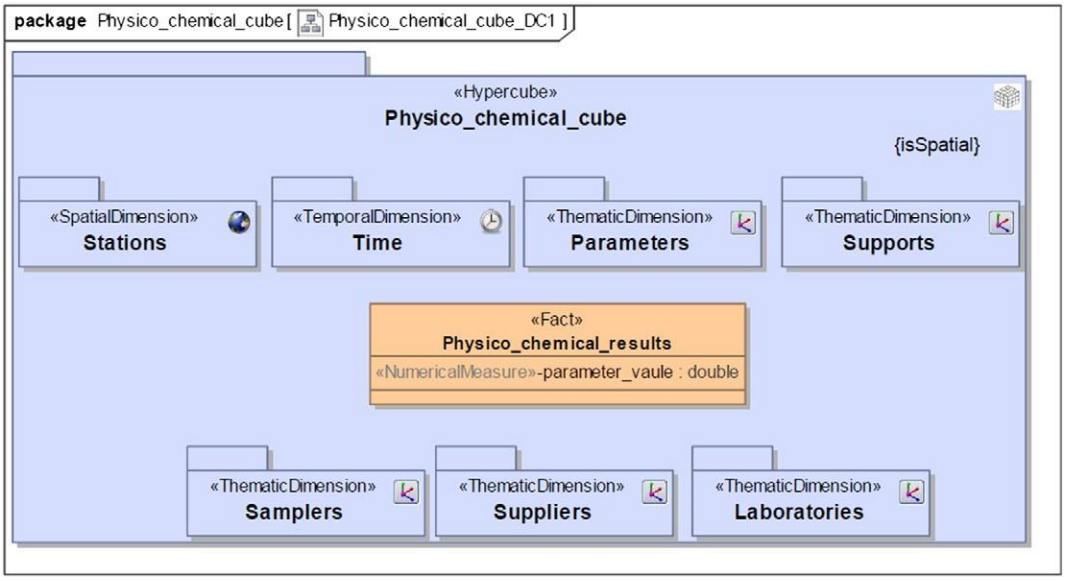

Fig. 4. The multidimensional model of the "physicochemical" data cube.

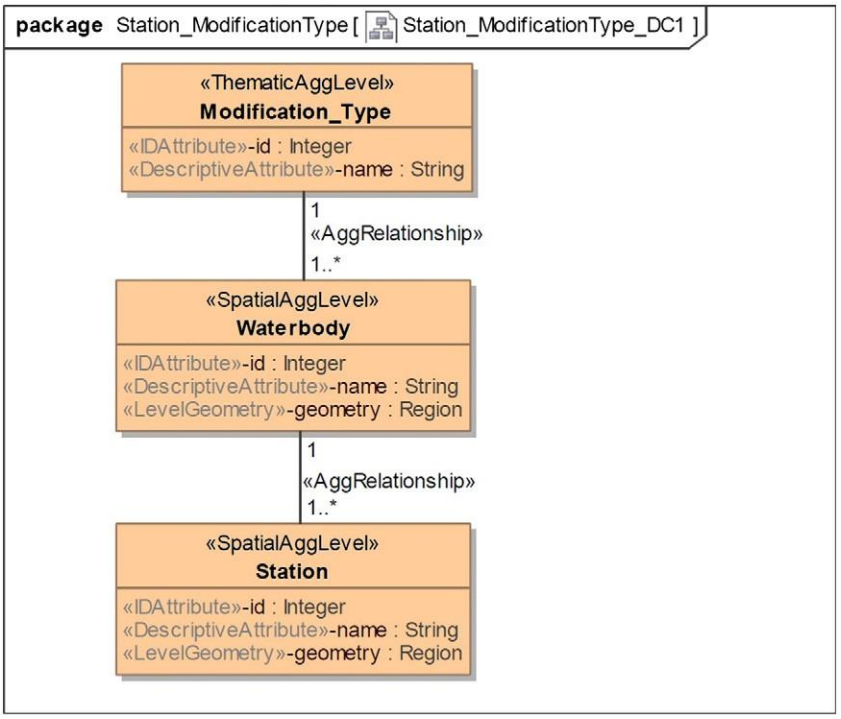

Fig. 5. An example of a spatial hierarchy.

This dimension is very important because it can prevent summing water quality measures of different supports.

(5) Sampler dimension. A thematic dimension describing the persons in charge of sampling.

(6) Laboratory dimension. A thematic dimension describing the laboratories in charge of sample data analysis to determine the water quality measure values. This dimension is important because laboratories use different analysis methods and tools and specific thresholding rules. Therefore, representing the laboratories as a dimension allows for pointing out these specificities by viewing water quality indicators by laboratory.

(7) Supplier dimension. A thematic dimension describing the organisms that order the samples at laboratories and supply external demanders with returned data.

According to Section 2, analysis indicators are calculated by aggregating measures using aggregate functions along hierarchies. For this cube, we defined several analysis indicators by applying different aggregate functions (Avg, Min, Max, etc.) to the measure "parameter value" (see Fig. 6). For example, the indicator "Average_parameter_values" is calculated by applying the average function (aggregator = 'Avg' of the aggregation rule) among all dimensions (the aggregation rule is of type AggRule, which means that it applies to all dimensions of the cube (see Section 2 for more details)). With the current cube modeling, the indicator Count_parameter_values yields only the number of parameter values (or fact table rows) for a combination of dimension members (e.g., a time period, a sub-set of physicochemical parameters and a sub-set of geographical zones). Based on this indicator, we will show in Section 5.3 how to define a more complex indicator, the presence count of physicochemical parameters, which is more informative and pertinent for water quality practitioners.

5.1.3. Hydrobiological data cube conceptual model

This cube allows for OLAP analysis of results of hydrobiological samples. The multidimensional diagram of this cube is depicted in Fig. 7. This model defines six biological water quality measures:

(1) the biological index score (e.g., for the French macrophyte index IBMR we have scores belonging to [0, 20]);

(2) the sample's abundance, i.e., the total number of individuals counted in one sample (e.g., 60 fishes of different species for the river fish index);

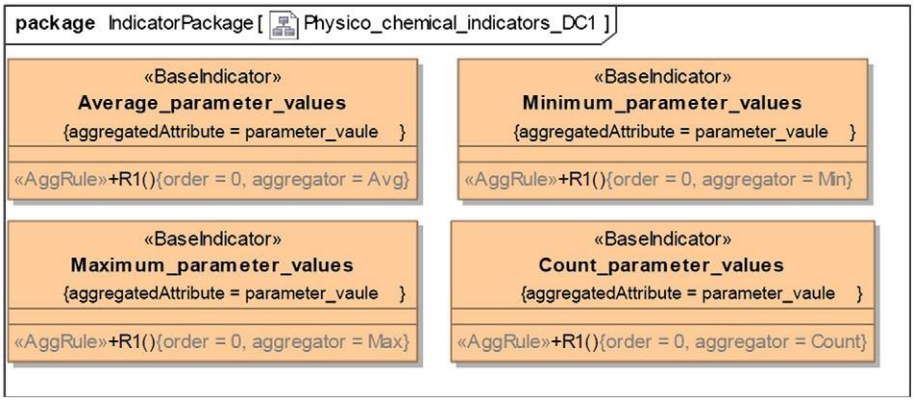

Fig. 6. The "physicochemical" simple analysis indicators.

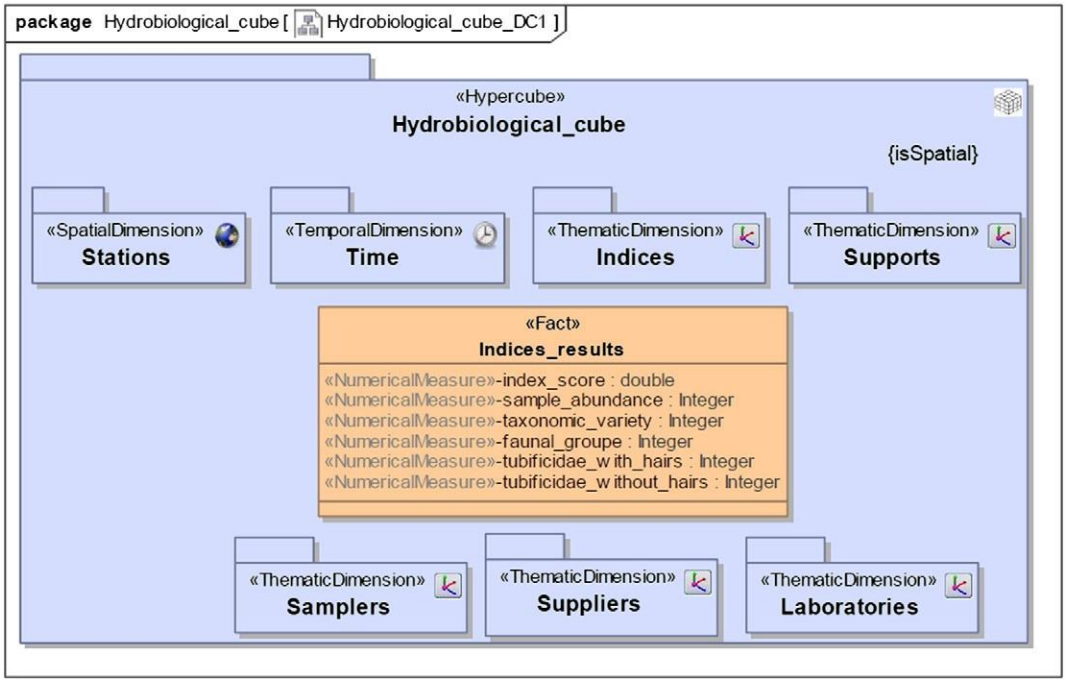

Fig. 7. The multidimensional model of the "hydrobiological" data cube.

(3) the taxonomic variety, the number of different species or taxa found in one sample for one index (e.g., 10 different types of plants for macrophytes);
(4) the faunal group for invertebrates, a value between 1 and 9, which corresponds to the most pollution-sensitive group of invertebrate families found in a sample of macroinvertebrates (e.g., 9 if the most pollution-sensitive families, Chloroperlidae and/or Perlidae, and/or Perlodidae and/or Taeniopterygidae, are found);
(5) the Tubificidae with hairs for oligochaetes indices, the number of individuals "Tubificidae with hairs" found in a sample of Oligochaeta (e.g., 7);
(6) the Tubificidae without hairs for oligochaetes indices, the number of individuals "Tubificidae without hairs" found in a sample of Oligochaeta (e.g., 30).

These measures are analyzed according to 7 dimensions (cf. Fig. 7). Six dimensions are identical to the dimensions of the physicochemical cube, one dimension is new.

(1) Indices dimension. A thematic dimension that groups biological indices (e.g., IBG, IBGA, IBGN, etc.) into taxonomic themes (e.g., invertebrates) to allow viewing of biological analysis indicators per indices and taxon theme.

For the OLAP analysis of the above-mentioned biological measures, we define many analysis indicators. For each measure, we define a certain number of indicators using different aggregate functions. Complex indicators and aggregate functions are described in Section 5.3. For now, we present only simple indicators that use common and OLAP tool-supported aggregate functions such as Avg, Min, Max, etc. For example, in Fig. 8, we show the simple indicators related to the "index_score" measure, each defined using a common aggregate function. For example, the "Minimum_index_score" is computed by applying the Minimum function to "index_ score" values. For now, the "Count_index_score" indicator gives only the number of scores (or fact table rows) for a combination of dimension members (e.g., a time period and a sub-set of indices). With the current cube modeling, this indicator does not give the count for a given index score (e.g., how many times we have the score 10 for the IBGN index). We will show in Section 5.3 how to make this indicator give this information without losing the current information possibility.

5.2. Implementation

To implement our solution, we chose a ROLAP architecture that is based only on free software tools (cf. Fig. 9).

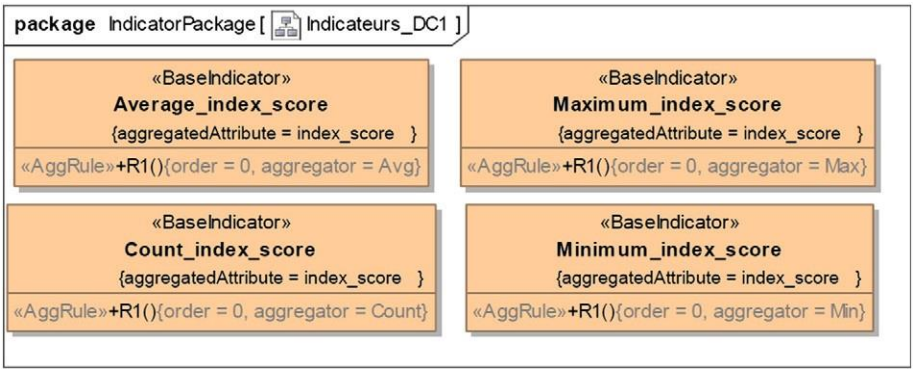

Fig. 8. The index score hydrobiological analysis indicators.

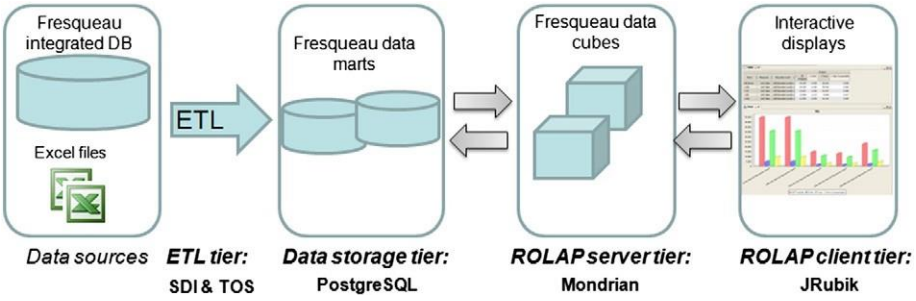

Fig. 9. Cube implementation architecture.

5.2.1. The ETL tier

The ETL tier allows for periodically populating the data marts with data from the data sources, which are in our application the integrated Fresqueau DataBase and some Excel files. The ETL tier consists of a set of data extraction and transformation JAVA programs implemented and performed using free ETL tools that are Spatial Data Integrator (SDI)[2] and Talend Open Studio (TOS).[3] Spatial Data Integrator is used particularly for the integration of spatial data.

5.2.2. The DW tier

The data storage tier or the project DW tier, which consists of two related[4] data marts (we have defined a data mart for each data cube), is implemented using a DataBase Management System (DBMS) PostgreSQL.[5] These data marts are defined following the star schema. As stated above (Section 2.3), this implementation schema defines one dimension table for each conceptual dimension. This schema type allows denormalized representations of dimensions. All of the aggregation levels of a dimension are stored in the same table. For example (Fig. 10), both levels of the "Indices" dimension ("Indices" and "Category") are mapped into the table "Indices", all levels of the "Stations" dimension into the "Stations" table, etc. Denormalization is used very often in DWs. This method produces redundancies (repetitions) of values but vastly improves the data access time. The star schema (as other DW schema types) uses a fact table to store measure values at the most detailed levels of dimensions. The fact table defines columns to represent conceptual measures and foreign keys that reference dimension tables to link these measure values to dimension data. For example, the fact table "Indices_results" stores index scores, sample abundances, taxonomic varieties, faunal groups, etc., by station, indices, day, support, sampler, laboratory, and supplier.

In terms of the sizes of the datasets, the total size of both data marts is approximately 4055 MB. The physicochemical fact table contains 14 602 580 rows, the hydrobiological fact table (Indices_results) contains 34 415 rows.

5.2.3. The OLAP server and client tiers

Additionally, we chose two other popular tools to explore and display data: Mondrian[6] as the OLAP server and JRubik[7] as the OLAP client. Mondrian is an open-source OLAP server that builds OLAP logical structures (e.g., dimensions, measures) on top of any DB on the basis of a specific XML file, called the Mondrian OLAP schema. This schema provides XML definitions for data cubes, dimensions, hierarchies, analysis indicators and their mappings to the DW/data mart data structures. In this schema, the XML definition of each dimension is between the XML elements bDimensionN and b/DimensionN, and every analysis indicator definition is between bMeasureN and b/MeasureN.

Finally, JRubik is a software package that provides a graphical presentation layer on top of Mondrian. This layer consists of a set of user-friendly interactive interfaces that trigger OLAP queries and display their results in different ways: pivot tables, statistical diagrams, maps, etc.

5.3. Modeling issues

In this section, we highlight some complex OLAP issues encountered in this project that are related mainly to the measure aggregation and the definition of the indicators, present the solutions proposed in the literature and show the practical solutions we adopted.

5.3.1. Complex indicators

In addition to hierarchies, the definition of analysis indicators is one of the fundamental parts of data cubes. As stated before in Section 3, analysis indicators, which can be simple or complex, are computed by aggregating measures using aggregate functions along hierarchies. A simple indicator involves a common aggregate function, a measure and all dimensions. Common aggregate functions are functions supported by DBMSs and OLAP tools (e.g., Sum, Avg and Count). A complex indicator can be defined as an application of different functions on different measures and along different dimensions, or as an application of a noncommon or complex aggregate function to a measure.

5.3.1.1. Complex aggregate functions. Complex aggregate functions are not supported by OLAP tools and need ad hoc definitions and developments by designers. In our project, examples of complex functions are standard deviation, percentile 10, median (percentile 50), percentile 90, and mode (the most frequent value). The standard deviation function is supported by the DBMS but not by the OLAP server Mondrian. The percentile P and mode functions are supported neither by the DBMS nor by the OLAP server.

To consider these functions in this application, we first added them to our conceptual framework (the UML profile) and then mapped them into implementations in ROLAP server and data storage tiers. Adding these functions to our UML profile allows designers to reuse them in other applications.

For example, to consider the functions percentile 1, …, and percentile 99, we extend our profile with the aggregate function (bbaggregatorNN) percentile (P) where 1 b = P b = 99, and propose implementations (i) in the data storage tier (PostgreSQL DW) in terms of PL/pgSQL stored procedures, and (ii) in the ROLAP server tier (Mondrian) in terms of MDX expressions. MDX (MultiDimensional eXpressions language) is a standard language to query multidimensional and OLAP databases, just like SQL for the relational databases. These implementations are easily transferable to other platforms (e.g., Oracle, MySQL, etc.). In Fig. 11, we show an example of usage of the percentile

[2] http://www.spatialdataintegrator.com.
[3] http://www.talend.com.
[4] The two data marts share some tables such as Suppliers, Laboratories and Samplers.
[5] http://www.postgresql.org.
[6] http://mondrian.pentaho.com.
[7] http://rubik.sourceforge.net/jrubik/intro.html.

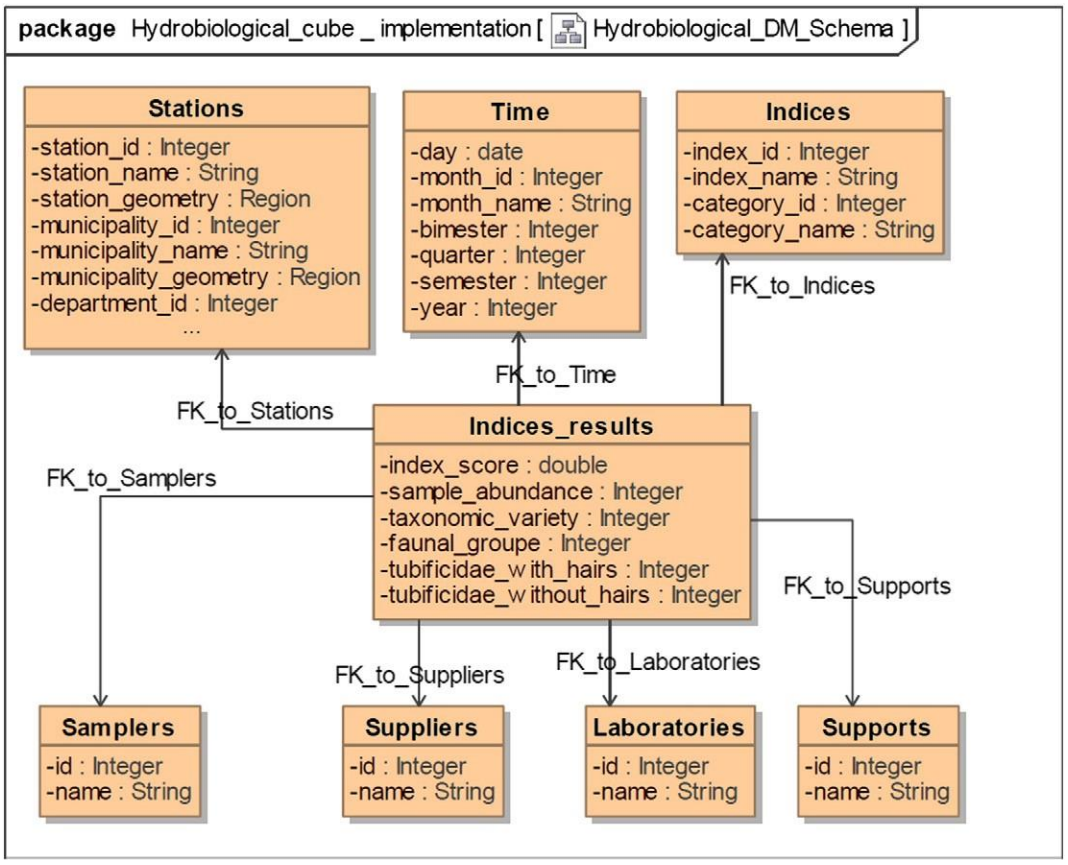

Fig. 10. The Indices data mart implementation schema.

function (added to our UML profile) to define three indicators, the percentiles 10, 50 (median) and 90 of physicochemical values.

The MDX implementation of the indicator "Percentile10_parameter_values" is:

This implementation uses a PL/pgSQL stored procedure "percentile_cont" that is implemented in PostgreSQL DBMS and allows for calculating all percentiles. The behavior of this procedure is identical to the PERCENTILE_COUNT Oracle function.[8]

5.3.1.2. Counting the presence/absence of parameters. Counting the presence/absence of physicochemical parameters such as pesticides is important to see, for example, how the appearance/disappearance of those parameters influences the water quality. In Section 5.1.2 (Fig. 6), we have defined a simple indicator, "Count_parameter_values", which gives only the number of parameter values for a given combination of dimension members by performing a nonconditional counting of fact table rows. To calculate the presence/absence of physicochemical parameters, we need to perform a conditional counting of physicochemical measure values by considering information concerning these values such as if they are in their domain of validity, below or above the detection threshold, etc. In the literature and existing OLAP tools, there is no aggregate function that allows such conditional counting.

To allow this conditional counting, we introduced an additional dimension that we call "Remarks". This dimension stores information about the validity and the exploitability of the results of the analysis of samples (e.g., in the domain of validity, below the detection threshold, etc.), allowing for conditional counting of values and, in general, for careful measure aggregations. This dimension is organized in a hierarchy of remark types (quantitatively exploitable, qualitatively exploitable, and nonexploitable) to ease the selection of its members by end users. Based on this dimension and the Count function, we defined the indicator "Count_of_Presence". Similarly to the problem of the heterogeneity of measurement units, we proposed two types of controls/implementations for this indicator: a non-automated control if we assume the end user awareness of the use of those remarks and an automated control based on an MDX implementation.

5.3.1.3. Counting index scores. Counting the index scores is useful for a definition of more interesting hydrobiological data analysis scenarios and a better interpretation of analysis results. Having the information of the count and distribution of scores over time periods, zones of station locations, etc., helps end users to formulate more pertinent analysis queries and interpret their results better. In Section 5.1.3, we have defined the Count_index_score indicator (Fig. 8) that gives only the count of score rows but not the information of distribution (how many times we have a given score or a score class for a given index and for a combination of other dimension members). To allow this distribution information, we defined an additional dimension, "Scores", which organizes scores in a hierarchy of 2 levels: score classes of level 2 are gathered into score classes of level 1. For example, the classes of level 2]0, 1],]1, 2], …,]9,10] are aggregated to the class of level 1] 0,10]. This hierarchy allows counting the number of times a score class occurs for an index and also facilitates selecting the Scores dimension members in the OLAP client.

This type of situation (a single attribute, index-score, having a dual usage, as a measure and as a dimension) is known in the literature as a degenerate dimension (Kimball and Ross, 2002; Luján-Mora et al., 2006). The solution proposed is to implement this type of attribute as a fact table column and to define a logical dimension that is mapped to this column (Kimball and Ross, 2002). Nevertheless, this solution does not allow for calculating the distributions (the number of times each value class of the measure occurs according to other dimensions).

[8] http://docs.oracle.com/cd/B19306_01/server.102/b14200/functions110.htm.

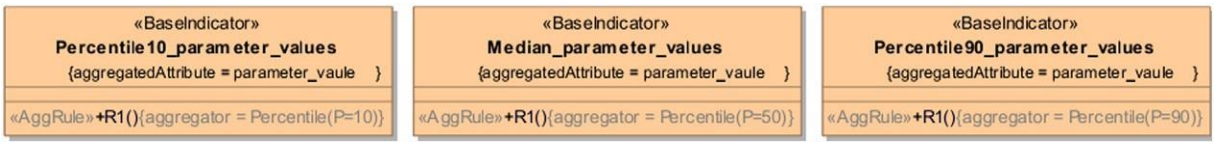

Fig. 11. Modeling of percentile indicators.

To overcome this limitation, we propose to implement this measure index-score as a fact table column and also to define a materialized dimension that is mapped to this column and records score classes.

5.3.2. Multiple heterogeneous measurement units

The correct aggregation of measures, known as summarizability, is a key issue in data cubes. To avoid incorrect indicator values, structural (such as the strictness of aggregation relationships—for example, a water quality station must be linked to one municipality in the Stations administrative hierarchy) and semantic conditions have to be verified by OLAP data structures/engines. The compatibility of measurement units is stated as one of the fundamental semantic conditions because aggregating measure values having different measurement units generally leads to meaningless results (e.g., summing micrograms per liter with grams per liter of physicochemical parameter values).

In the Fresqueau project, physicochemical results are provided by data suppliers with multiple heterogeneous measurement units (microgram per liter, centimeter per minute, cubic centimeter, gram, gram per meter squared, etc.). This heterogeneity is essentially because each laboratory has its own objectives and management rules as well as different objectives and conventions for the source databases.

To address this problem of measurement unit heterogeneity and allow correct and meaningful aggregate values, we use two solutions.

(1) Define the "Measurement units" as a dimension of the physicochemical cube. This definition is very important to control the measure aggregation (prevent summing results with incompatible measurement units, for example, grams with liters) and also for information rendering aspects (displaying measurement units is necessary for end users to understand and to interpret results). This solution can be applied independently of the convertibility between measures. In the OLAP client, this control can be automated or not. For the nonautomated control, the user can select the compatible measures by selecting them in coordinates of the pivot table (if he/she wants to visualize them) or in the filter zone to focus the aggregation on them. To fasten this selection/filtering, we can define a hierarchy for this dimension by grouping measurement units into categories according to their compatibility. The automated control is achieved by adding some test conditions (that are expressed in terms of MDX expressions) to the definitions of analysis indicators.

(2) ETL transformations. This solution is implemented in the ETL tier. The solution consists of transforming the results having convertible measurement units into the same unit. For example, by converting parameter values having g/L and mg/L into µg/L by multiplying by 10^6 and by 10^3, respectively. The first solution is needed because we have many nonconvertible measurement units (e.g., L, kg, mg, L/cm^2, etc.).

6. OLAP analysis

OLAP analysis is generally performed in an exploratory way following a top-down approach (Sapia, 1999). In an OLAP analysis session which is a sequence of queries, the user starts by formulating a first coarser query by selecting (on the OLAP client interface) the indicators and dimensions level (that he/she wants to display in the result) at their most coarse granularity levels. Based on this result, the user can perform further finer analysis by clicking on interface components (buttons, maps, etc.) that trigger OLAP operations, or by formulating other combinations of dimensional elements.

In this section, we present examples of different types of OLAP analysis, thematic, temporal, spatial, spatiotemporal and multiscale analysis, using the hydrobiological data cube, to show the feasibility and the productivity which can be gained by end users by using OLAP solutions in the field of water quality assessment. We also illustrate three types of visualizations (pivot table, pie charts and histogram diagrams). Obviously, many other basic and advanced operations and visualizations are available allowing for finer and more complex analysis.

6.1. Thematic multiscale analysis

Thematic multiscale analyses are performed using/along thematic dimension hierarchies such as "Supports", "Suppliers", "Samplers" and "Laboratories". As stated before, OLAP analyses are generally performed in a top-down way, going from larger queries (summaries) to finer queries. Thus, the first queries should show indicator values at the coarsest aggregation levels of dimensions. Next, we show examples of thematic queries.

Query 1. This query shows the average and standard deviation values per index and for all laboratories and all samplers. To increase the readability, only five indices are represented, one for each taxonomic theme: the Specific Pollution-Sensibility Index of diatoms ("IPS" in French), the Normalized Global Biological Index of invertebrates ("IBGN"), the standardized River Macrophyte Biological Index ("IBMR"), the Oligochaetes Sediment Bioindication Index ("IOBS"), and the River Fish Index ("IPR"). IBGN, IBMR, and IPS values are positive real numbers in the range of [0, 20]; IOBS values are in the range of [0, 10]; and IPR values are in the range of [0, ∞]. For the IBGN, IBMR and IPS, the best score is 20; the IOBS best score is 10 and the IPR best score is 0. Fig. 12(a) shows a pivot table representation of the results, and Fig. 12(b) shows a histogram diagram representation. With these representations, we can rapidly see, for example, that the IOBS is particularly low, and an expert can conclude that for this dataset, pollution-sensitive oligochaetes have disappeared and sediments most likely are in a bad state. The statistical diagram representation is synchronized with the pivot table representation: every change in the pivot table representation is instantaneously reproduced in the diagrammatic representation. In terms of query response time, the results are displayed instantly after the execution of the query by the end user.

Query 2. To illustrate a multiscale thematic analysis, the query represented in Fig. 13 shows the same indicators as above, the average and the standard deviation values of index scores, at the "Supplier" aggregation level (thematic scale), by index and for all laboratories. The results of this query are obtained from the query 1 by performing a Drill-down OLAP operation on the "Suppliers" dimension hierarchy. The drill-down operation as shown here increases the level of detail of indicators. For more readability, we show only the "Average_index_score" indicator values. In the pivot table of Fig. 13 and inversely to the pivot table above (Fig. 12(a)), the laboratories, samplers and indicators are represented in rows, and the indices in columns. The pivoting of the table axis is performed instantly (after the drilling down operation) using the Rotate OLAP operator.

6.2. Temporal multiscale analysis

Temporal multiscale analyses are possible through different temporal aggregation levels contained in both temporal hierarchies

a

	Laboratories_hierarchy
	all laboratories
	Samplers_hierarchy
	all samplers
Indices_hierarchy	Average_index_score / StDeviation_index_score
IPS	13 / 2,243
IBGN	13 / 3,709
IBMR	8 / 2,385
IOBS	3 / 10,516
IPR	18 / 13,281

b

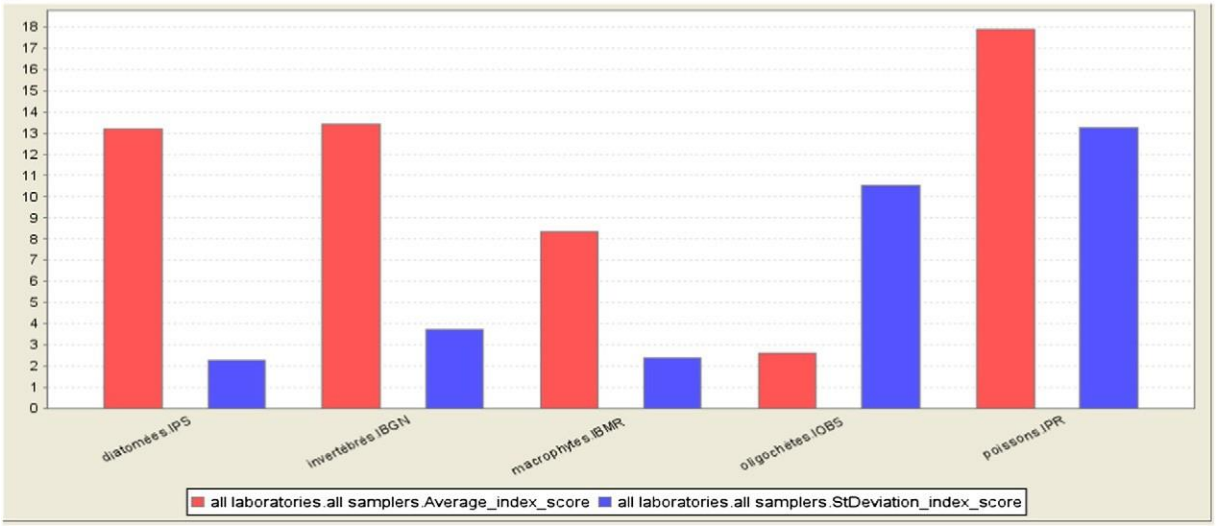

Fig. 12. The average and standard deviation of index scores by index and for all samplers and all laboratories. (a) Pivot table visualization. (b) Histogram visualization.

(cf. Section 5.1.2). An example of a temporal query is described next.

Query 3 (Fig. 14). This query represents the average index scores (the "Average_index_score" indicator values), by index, by year (a temporal scale), and for all samplers. The results are displayed with a pivot table representation in Fig. 14(a) and a histogram representation in Fig. 14(b). For more readability, we consider only the years between 2000 and 2010. Fig. 14(a) shows a visualization of results using a pivot table where the dimensions "Indices" and "Suppliers" ("all suppliers" member) are placed in columns and the "Time" dimension in rows. Fig. 14(b) shows a pie chart diagram representation of these results (a diagram per index), which allows a rapid viewing of the distribution of scores of every index over the time period considered (2000 to 2010).

6.3. Spatial multiscale analysis

Spatial multiscale analyses are rendered possible through the different spatial granularities of the "Stations" dimension hierarchies (cf. Section 5.1.2). An example of a spatial query is shown next.

Query 4 (Fig. 15). This query shows the average index scores by index, for all samplers, at the hydroecoregion spatial scale. The Indices and Samplers dimensions with the indicator "Average_index_score"

			Indices_hierarchy				
Laboratories_hierarchy	Samplers_hierarchy	Mesures	IPS	IBGN	IBMR	IOBS	IPR
all laboratories	Agence de l'Eau Adour-Garonne	Average_index_score		15			
	Agence de l'Eau Artois-Picardie	Average_index_score		15			
	Agence de l'Eau Loire-Bretagne	Average_index_score		14			
	Agence de l'Eau Rhin-Meuse	Average_index_score		14			
	Agence de l'Eau Seine-Normandie	Average_index_score		16			
	Laboratoire Départemental d'Analyses de l'Allier	Average_index_score		13			
	Preleveur inconnu	Average_index_score	13	13	8	3	18
	Service de Bassin Loire-Bretagne	Average_index_score		18			
	Service Hydrologique Centralisateur Adour	Average_index_score		13			
	Service Hydrologique Centralisateur Garonne	Average_index_score		15			

Fig. 13. The average of index scores by index, by sampler and for all laboratories.

a

	Samplers_hierarchy				
	all samplers				
	Indices_hierarchy				
	IPS	IBGN	IBMR	IOBS	IPR
Time_hierarchy_1	Average_index_score	Average_index_score	Average_index_score	Average_index_score	Average_index_score
2000		13	9		
2001		14	9		18
2002	13	13	8		18
2003	13	14	8		19
2004	11	13	8		18
2005		14	8	1	17
2006	16	14	9	1	17
2007	13	15	9	1	18
2008	13	14	9	1	19
2009	13	14			18
2010	13	15			17

b

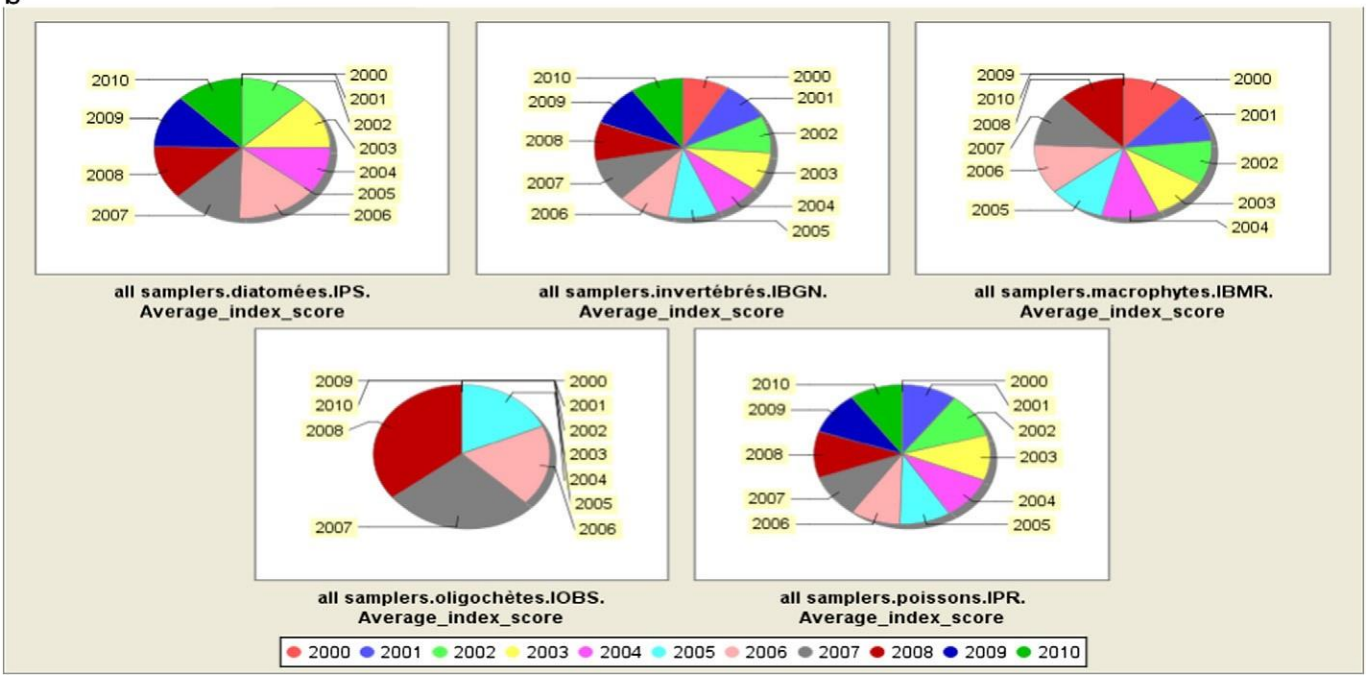

Fig. 14. The average of index scores by index, by year and for all samplers. (a) Pivot table visualization. (b) Histogram visualization.

are placed in columns and the spatial dimension hierarchy, "Hydroecoregion_hierarchy", in rows.

6.4. Spatiotemporal multiscale analysis

Hybrid analyses can be performed by combining thematic, temporal, and spatial dimension hierarchies when exploring the data. Spatiotemporal analyses in particular allow for viewing indicator values at various combinations of spatial and temporal scales of the defined Stations and Time dimensions. Examples of spatiotemporal analysis are shown below.

Query 5 (Fig. 16). This query shows the average index scores by index, year (the temporal scale) and for all hydroecoregions (the spatial scale). The Indices and Spatial dimension hierarchies are put in columns and the temporal hierarchy in rows. The results are displayed with a pivot table representation in Fig. 16(a) and with a histogram representation in Fig. 16(b).

Query 6 (Fig. 17). This query shows the average index scores by index, hydroecoregion and year. Query 6 is obtained from query 5, after a drill-down operation on the spatial member "all hydroecoregions". The table axes of query 6 are also pivoted (using the Rotate operation) for more readability.

We can choose other hydrobiological indicators (such as "Minimum_index_score" and "Maximum_index_score"), display many indicators and dimensions at the same time, and use different diagrammatic visualizations, etc.

7. Conclusions and future work

In this paper, we have shown an application of the OLAP technology to the field of water quality assessment. Based on our framework for data cubes (Boulil et al., 2013a), we developed a free tool-based and extensible ROLAP system composed of two data cubes: (1) a data cube for the OLAP analysis of physicochemical water quality data, and (2) a data cube for the OLAP analysis of hydrobiological data. We proposed standards-based (UML, SQL and MDX) and generic solutions to model and implement complex indicators using complex aggregates such as percentiles by extending our framework with these functions. Other complex analyses are defined by introducing useful dimensions and using common aggregate functions. Additionally, we proposed two

	Samplers_hierarchy				
	+ all samplers				
	Indices_hierarchy				
	IPS	IBGN	IBMR	IOBS	IPR
Hydroecoregion_hierarchy	Average_index_score	Average_index_score	Average_index_score	Average_index_score	Average_index_score
−all hydroecoregions	13	13	8	3	18
+ALPES INTERNES		12			24
+ALSACE	13	10			17
+ARDENNES	13	13			14
+CEVENNES		16			12
+CORSE		16			15
+COTEAUX AQUITAINS		14			53
+COTES CALCAIRES EST	13	13			23
+HER1 inconnue	16	12	8	3	17
+JURA-PREALPES DU NORD	12	13			17
+MASSIF CENTRAL SUD		13			19
+MEDITERRANEEN		13			22
+PLAINE SAONE		13			25
+PREALPES DU SUD		14			20
+PYRENEES		16			17
+VOSGES	15	14			13

Fig. 15. The average of index scores by index, for all samplers and by hydroecoregion.

practical solutions to address the summarizability problem of heterogeneous measurement units. Finally, to validate our system and to show the productivity that can be gained by water quality practitioners when using OLAP, we provided the reader with a number of examples of OLAP analysis.

This work has been achieved in the context of the Fresqueau project which aims to develop new methods for collecting, analyzing and interpreting all available data related to water quality parameters. The built data cubes proved to be relevant and usable tools to help domain experts (mainly hydroecologists) exploring, selecting and analyzing the huge and complex datasets collected in the Fresqueau database. These users are very satisfied with the rapidity, interactivity and facility of the analyses that can be performed using the developed OLAP system, the variety of results according to the

a

	Hydroecoregion_hierarchy				
	+ all hydroecoregions				
	Indices_hierarchy				
	IPS	IBGN	IBMR	IOBS	IPR
Time_hierarchy_1	Average_index_score	Average_index_score	Average_index_score	Average_index_score	Average_index_score
+2000		13	9		
+2001		14	9		18
+2002	13	13	8		18
+2003	13	14	8		19
+2004	11	13	8		18
+2005		14	8	1	17
+2006	16	14	9	1	17
+2007	13	15	9	1	18
+2008	13	14	9	1	19
+2009	13	14			18
+2010	13	15			17

b

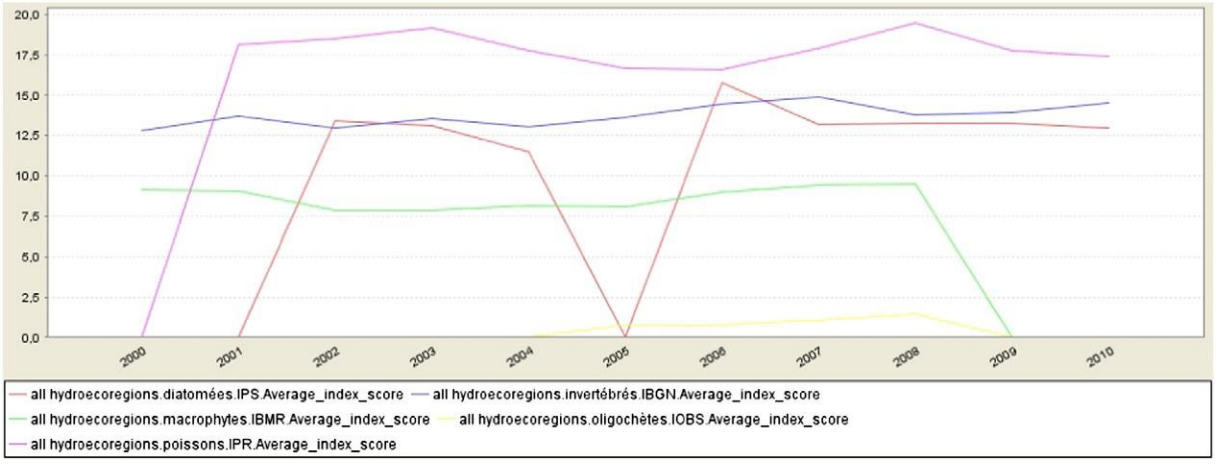

Fig. 16. The average index scores by year, index and for all hydroecoregions. (a) Pivot table visualization. (b) Histogram visualization.

			Time_hierarchy_1										
Hydroecoregion_hierarchy	Indices_hierarchy	Mesures	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
ALPES INTERNES	IBGN	Average_index_score	12	13	11	13	12	12	13	13	12	13	13
ALSACE	IBGN	Average_index_score			9	9	9	11	9	11	13	14	14
ARDENNES	IBGN	Average_index_score			14	14	17	13	10				
CEVENNES	IBGN	Average_index_score	14	17	14	18	13	14	18	18	17	17	18
CORSE	IBGN	Average_index_score	16	17	16	16	13		16	12	15	14	14
COTEAUX AQUITAINS	IBGN	Average_index_score	16	16	10		14		15	15	15	15	15
COTES CALCAIRES EST	IBGN	Average_index_score	12	12	12	15	14	13	14	16	15	15	15
HER 1 inconnue	IBGN	Average_index_score	11	10	11	15	12	13	12	13	13	11	14
JURA-PREALPES DU NORD	IBGN	Average_index_score	13	13	13	13	15	16	16	15	13	14	15
MASSIF CENTRAL SUD	IBGN	Average_index_score	13	12	13	14	13	15	15	15	13	14	14
MEDITERRANEEN	IBGN	Average_index_score	13	13	13	13	12	13	14	15	13	14	14
PLAINE SAONE	IBGN	Average_index_score	12	13	14	13	14	15	16	15	13	14	14
PREALPES DU SUD	IBGN	Average_index_score	14	13	15	14	14	15	15	16	15	15	15
PYRENEES	IBGN	Average_index_score	16	16	16		17	14	17	18	16	17	17
VOSGES	IBGN	Average_index_score	10	12	13	15	13	14	14	16	16	15	16

Fig. 17. The average index scores by index, hydroecoregion and year.

various aggregation levels, and the new viewpoints on data they can thus obtain.

The next steps in this work are to extend/enrich our ROLAP system by: (1) other data cubes (such as a data cube for hydromorphological data of water bodies, a data cube for environmental forcing variables such as flows, etc.), (2) linking the data cubes to each other to allow drill-across OLAP operations, to answer the questions related to interdependencies between the water quality parameters such as the influence of physicochemical state of water bodies in the appearance/disappearance of faunal and floristic specifies, the influence of hydromorphological characteristics of water bodies in their physicochemical states, etc. Another perspective is to study the possibilities offered by spatial OLAP tools (Bédard et al., 2007; Miquel et al., 2010). These tools may help end users to understand water quality data better by allowing map visualizations and explorations of data. Finally, we also plan to develop an architecture that integrates or connects the OLAP tool to a data mining module consisting of different data mining algorithms (Wang and Guo, 2013). The OLAP module will be used for a first and rapid exploration of the data. Based on the OLAP exploration results, the end user can execute an adequate algorithm of the data mining module on the adequate dataset to discover additional knowledge such as existence or absence of correlations between water quality parameters, etc.

Acknowledgments

This work has been funded by the National Research Agency (ANR), as a part of the ANR11 MONU14 Fresqueau project.

References

Abelló, A., Samos, J., Saltor, F., 2006. A multidimensional conceptual model extending {UML}. Inf. Syst. 541–567. http://dx.doi.org/10.1016/j.is.2004.12.002.

AFNOR, 2004. Qualité de l'eau - détermination de l'Indice Biologique Global Normalisé (IBGN), XP T90-350.

AFNOR, 2009. Qualité de l'eau - Prélèvement des macro-invertébrés aquatiques en rivières peu profondes, XP T90-333.

AFNOR, 2010. Qualité de l'eau - Traitement au laboratoire d'échantillons contenant des macro-invertébrés de cours d'eau, XP T90-3888.

Alexandru, A., Gorghiu, G., Leane Carmen, N., Alexandru, C.-A., 2010. Using OLAP systems to manage environmental risks in Dambovita County. Bull. UASVM Hortic. 67 (2), 394–399.

Bédard, Y., Gosselin, P., Rivest, S., Proulx, M.-J., Nadeau, M., Lebel, G., Gagnon, M.-F., 2003. Integrating gis components with knowledge discovery technology for environmental health decision support. Int. J. Med. Inform. 70, 79–94. http://dx.doi.org/10.1016/S1386-5056(02)00126-.

Bédard, Y., Rivest, S., Proulx, M.-J., 2007. Spatial on-line analytical processing (SOLAP): concepts, architectures and solutions from a geomatics engineering perspective. In: Wrembel, R., Koncillia, C. (Eds.), Data warehouses and OLAP: concepts, architectures and solutions. IRM Press, pp. 298–319 (Chap XIII).

Boulil, K., Bimonte, S., Pinet, F., 2012. A UML {&} spatial OCL based approach for handling quality issues in SOLAP systems. ICEIS, 1, pp. 99–104.

Boulil, K., Bimonte, S., Pinet, F., 2013a. Conceptual model for spatial data cubes: a UML profile and its automatic implementation. Comput. Stand. Interfaces. http://dx.doi.org/10.1016/j.csi.2014.06.004.

Boulil, K., Pinet, F., Bimonte, S., Carluer, N., Lauvernet, C., Cheviron, B., Miralles, A., Chanet, J.-P., 2013b. Guaranteeing the quality of multidimensional analysis in data warehouses of simulation results: application to pesticide transfer data produced by the {MACRO} model. Ecol. Inform. 16, 41–52. http://dx.doi.org/10.1016/j.ecoinf.2013.04.004.

Chen, Z., Gangopadhyay, A., Holden, S.H., Karabatis, G., McGuire, M.P., 2007. Semantic integration of government data for water quality management. Gov. Inf. Q. 24, 716–735. http://dx.doi.org/10.1016/j.giq.2007.04.004.

Codd, E.F., Codd, S.B., Salley, C.T., 1993. Providing OLAP (on-line analytical processing) to user-analysts: an IT mandate Codd Date 32.

Glorio, O., Trujillo, J., 2008. An MDA approach for the development of spatial data warehouses. Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery. Springer-Verlag, Berlin, Heidelberg, pp. 23–32. http://dx.doi.org/10.1007/978-3-540-85836-2_3.

Hahn, K., Sapia, C., Blaschka, M., 2000. Automatically generating OLAP schemata from conceptual graphical models. Proceedings of the 3rd ACM International Workshop on Data Warehousing and OLAP. ACM, New York, NY, USA, pp. 9–16. http://dx.doi.org/10.1145/355068.355310.

Inmon, W.H., 2005. Building the data warehouse. Wiley. com.

Kimball, R., Ross, M., 2002. The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling. John Wiley & Sons, Canada.

Lalande, N., Berrahou, L., Molla, G., Serrano, E., Cernesson, F., Grac, C., Herrmann, A., Le Ber, F., Teisseire, M., Trémolières, M., 2013. Feedbacks on data collection, data modelling and data integration of large datasets: application to Rhin-Meuse and Rhone-Mediterranean districts. 8th Symposium for European Freshwater Sciences. Münster, Germany, p. Oral communication of 30 minutes.

Luján-Mora, S., Trujillo, J., Song, I.-Y., 2006. A UML profile for multidimensional modeling in data warehouses. Data Knowl. Eng. 59, 725–769. http://dx.doi.org/10.1016/j.datak.2005.11.004.

Mahboubi, H., Bimonte, S., Deffuant, G., Chanet, J.-P., Pinet, F., 2013. Semi-automatic design of spatial data cubes from simulation model results. IJDWM 9, 70–95.

Malinowski, E., Zimányi, E., 2008. Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications; with 10 Tables. Springer.

McGuire, M.P., Gangopadhyay, A., 2006. Modeling, Visualizing and Mining Hydrologic Spatial Hierarchies for Water Quality Management.

McGuire, M., Gangopadhyay, A., Komlodi, A., Swan, C., 2008. A user-centered design for a spatial data warehouse for data exploration in environmental research. Ecol. Inform 3, 273–285. http://dx.doi.org/10.1016/j.ecoinf.2008.08.002.

Miquel, M., Bimonte, S., Pinet, F., Tchounikine, A., 2010. When spatial analysis meets OLAP: multidimensional model and operators. Int. J. Data Warehous. Min. 6, 33–60. http://dx.doi.org/10.4018/jdwm.2010100103.

OMG, 2011. Unified Modeling LanguageTM (OMG UML), Infrastructure, Version 2.4.

Pardillo, J., Mazón, J.-N., 2010. Designing OLAP schemata for data warehouses from conceptual models with MDA. Decis. Support. Syst. http://dx.doi.org/10.1016/j.dss.2010.04.006.

Pinet, F., Schneider, M., 2010. Precise design of environmental data warehouses. Oper. Res. 10, 349–369.

Radulescu, C.Z., Radulescu, M., 2008. A Multidimensional Data Model for Environment Protection. Proceedings of the 12th WSEAS International Conference on Computers. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, pp. 1101–1106.

Romero, O., Abelló, A., 2009. A survey of multidimensional modeling methodologies. Int. J. Data Warehous. Min. 5, 1–23.

Sapia, C., 1999. On modeling and predicting query behavior in OLAP systems. DMDW, p. 2.

Scotch, M., Parmanto, B., 2006. Development of SOVAT: a numerical–spatial decision support system for community health assessment research. Int. J. Med. Inform 75, 771–784. http://dx.doi.org/10.1016/j.ijmedinf.2005.10.008.

Shekhar, S., Lu, C.T., Liu, R., Zhou, C., 2002. CubeView: a system for traffic data visualization. Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference On, pp. 674–678. http://dx.doi.org/10.1109/ITSC.2002.1041299.

Stefanovic, N., Han, J., Koperski, K., 2000. Object-based selective materialization for efficient implementation of spatial data cubes. Knowl. Data Eng. IEEE Trans. 12, 938–958. http://dx.doi.org/10.1109/69.895803.

Strahler, A.N., 1957. Quantitative analysis of watershed geomorphology. Trans. Am. Geophys. Union 38, 913–920.

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

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. http://dx.doi.org/10.1016/j.agsy.2013.07.005.

Wang, H.C., Guo, J.-L., 2013. Constructing a water quality 2.0 OLAP system in Taiwan. J. Clean. Prod 40, 40–45. http://dx.doi.org/10.1016/j.jclepro.2011.04.019.