Data Science approach for a cross-disciplinary understanding of urban phenomena: Application to energy efficiency of buildings

Sylvie Servigne, Yann Gripay, Jean-Michel Deleuil, Céline Nguyen, Jacques Jay, Olivier Cavadenti, Mebrouk Radouane

Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 115 (2015) 45 – 52

Procedia Engineering

www.elsevier.com/locate/procedia

Symposium "Towards integrated modelling of urban systems"

Data Science approach for a cross-disciplinary understanding of urban phenomena: Application to energy efficiency of buildings

Servigne Sylvie[1], Gripay Yann[1], Deleuil Jean-Michel[2], Nguyen Céline[2], Jay Jacques[3], Cavadenti Olivier[1], Mebrouk Radouane[1]

[1]*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, Labex IMU, Laboratoire d'Informatique en Image et Systèmes d'Information,*
[2]*Université de Lyon, CNRS, INSA-Lyon, EVS, UMR5600, Laboratoire Environnement, Ville et Société*
[3]*Université de Lyon, CNRS, INSA-Lyon, CETHIL, UMR5008, Centre de Thermique de Lyon*
F-69621 Villeurbanne

Abstract

Our goal is to develop theoretical and practical tools to model, explore and exploit heterogeneous data from various sources in order to understand a phenomenon. We focus on a generic model for data acquisition campaigns based on the concept of generic sensor. The concept of generic sensor is centered on acquired data and on their inherent multi-dimensional structure, to support complex domain-specific or field-oriented analysis processes. We consider that a methodological breakthrough, based on Data Science as a pivot for interdisciplinary dialog, may pave the way to deep understanding of voluminous and heterogeneous scientific data sets. Our use case concerns energy efficiency of buildings to understand the relationship between physical phenomena and user behaviors. This multidisciplinary project involves computer scientists, social and urban scientists, and thermal scientists. The aim of this paper is to give a synthetic presentation of our methodology, and an overview of our main results.

Peer-review under responsibility of LET

Keywords: Sensors – Data Model - Heterogenous Data – Multidimensional Data - Data Correlation – Data Exploration

1. Context and motivation

Urban-related phenomena involve numerous contextual interactions between humans, infrastructures, objects, etc. In order to study such complex phenomena, researchers and institutions have multiple ways to collect data through dedicated campaigns: monitoring of infrastructures like roads, buildings or urban networks; environmental monitoring for air, water, soil, and so on; surveys on users, inhabitants and citizens; historic and sociological studies... Those campaigns can generate a large amount of heterogeneous and complex data, and those data are multi-sources, multi-dimensional (e.g., spatiotemporal data), and multimedia (e.g., numbers, texts, images, sounds, videos).

Peer-review under responsibility of LET
doi:10.1016/j.proeng.2015.07.353

To achieve the goal of hypothesis validation, and eventually knowledge discovery, those data needs to be deeply analyzed by multidisciplinary teams. However, deep analysis requires advanced skills first to understand raw data, then to discover their multi-scale properties, and finally to perform relevant aggregations and cross-source comparisons.

In our project, we focus on the context of building-related phenomena. Nowadays, energy efficiency of a building is given by theoretical estimations based on its initial design and on typical usage scenarios. However, the practical efficiency is usually lower due to the complexity of the building process and to the real behavior of occupants. If energy consumption and environmental conditions can be directly measured through instrumentation, understanding the practical energy efficiency of a building requires a multidisciplinary approach, in order to understand energy consumption with regard to actual uses of the building. Cross-analyses of "instrumentation data" and "survey data" (and other studies data) are thus necessary to fully discover and then understand complex correlations between physical and human parameters.

Our objective is to develop theoretical and practical tools to model, explore and exploit heterogeneous data from various sources in order to understand a phenomenon. We focus on a generic model for data acquisition campaigns based on the concept of generic sensor. The concept of generic sensor is centered on acquired data and on their inherent multi-dimensional structure. This multi-dimensional structure is then a support for complex domain-specific or field-oriented analysis processes. We consider that a methodological breakthrough, based on Data Science as a pivot for interdisciplinary dialog, may pave the way to deep understanding of voluminous and heterogeneous scientific data sets.

We first present our methodology for a cross-disciplinary understanding of urban-related phenomena, in Section 2. We then discuss about related works in Section 3. In Section 4, we give an overview of our main results concerning data model, exploration language and visualizations.

2. Methodology

Our approach revolves around a generic conceptual model of sensor data. This model has been built from a real multi-disciplinary approach, named Data Science approach that involves researchers from computer science, human science and thermal science. A generic model would permit the representation of heterogeneous data through only a few generic concepts.

In order to build this final model, we first took a bottom-up approach. We designed a preliminary model for heterogeneous **physical sensors**, based on a real experimental platform in occupied buildings. We designed another preliminary model for **sociological surveys**, to take into account opinions and feelings of occupants "measured" by a questionnaire.

We then took a top-down strategy. Inspired by existing abstract ontologies that describe sensor systems [1, 2, 3], we designed the **Virtual Generic Sensor model** (VGS model) that can homogeneously represents data from heterogeneous physical sensors, data from sociological surveys, and data from other kinds of sources. This model is also based on a previous sensor model designed for natural risk monitoring [4, 5, 6]. The VGS model focuses on **data** produced by **generic sensors** linked to a **common multi-dimensional structure**. This multi-dimensional structure describes time, location and source of measured data, and is designed to support additional specific field-oriented dimensions.

We also designed a methodology for an **agile multi-dimensional exploration** of those data. Based on the VGS model and its multi-dimensional structure, we propose a language to finely define domain-specific or field-oriented indicators through successive aggregations along dimensions [12] (in a similar way to Data Warehouses). We designed a **visualization framework** linked to those dimensions that enables users to visually explore indicators using graphs. And in order to visually compare indicators, we propose an interactive "matrix layout" for those graphs. The common multidimensional structure of data is then exploited at three levels: to structure data, to define indicators, and to explore data with these indicators.

Our agile approach allows **incremental** and **iterative** data processes and analysis. Users can start to explore raw data with some predefined dimensions for time, source and location of measured data, and with basic aggregated indicators like MIN, MAX, AVG. Exploration can then be incrementally and iteratively enriched by users themselves. At the data level, we consider incremental data sets, e.g., when raw data are still being captured and appended to the data set. The data set generation is also iterative: data can be progressively enriched with new interpreted data that may then be used in the same fashion as raw data. At the analysis level, we consider an incremental exploration process: new (aggregated) indicators can be added on-the-fly, when needed by users. The exploration process is also iterative when knowledge from past explorations is used to refine current and future

explorations: existing analysis dimensions can be refined, with more precise levels of granularity in their level hierarchy, and new dimensions can even be added, in order to offer new points of view on data.

This approach is designed to support and enrich current domain-specific approaches for complex and/or scientific data analysis. In particular, each visualization of data is precisely and concisely described by its "lineage", i.e. by the link to the data subset, data aggregation definitions, and visual projection parameters that lead to this data visualization.

3. Related works

Sensors have started becoming an integral part of our personal lives, for example in the form of temperature and humidity sensors, or smoke and fire detectors. Examples of some wireless sensors are shown in Figure 3. Sensors can send periodical measurements for long periods, with only very little human intervention. Depending on the chosen periodicity of data acquisition, these sensors can produce a large amount of data in a very short amount of time.

Sensor data modelling has recently generated a number of research works. Many sensor ontologies have been proposed since 2005.They may focus on the description of the observation process, on the structure of the sensor network, or on the description of the physical sensors. They often contain specificities from a targeted application domain.

Towards an objective of standardization, the W3C incubator group SSN XG3 proposed the Semantic Sensor Network (SSN) ontology based on SensorML [13] and Dolce UltraLite [2]. SSN ontology describes key concepts related to sensors and sensor networks. It proposes a collection of models for representing the main concepts linked to the context of sensors and defining the relations between these concepts, like Process, Stimulus-Sensor-Observation Pattern, Device, Data. It allows a separation between hardware aspects and abstract processes of a sensor. SSN therefore offers an ontology adapted to the description of a sensing system. However SSN descriptions are abstract and focus on physical sensors and sensing systems. The specification of abstract concepts for data measures is less detailed, with only one concept of Observation Value in the Data model (at the bottom left of Figure 1).

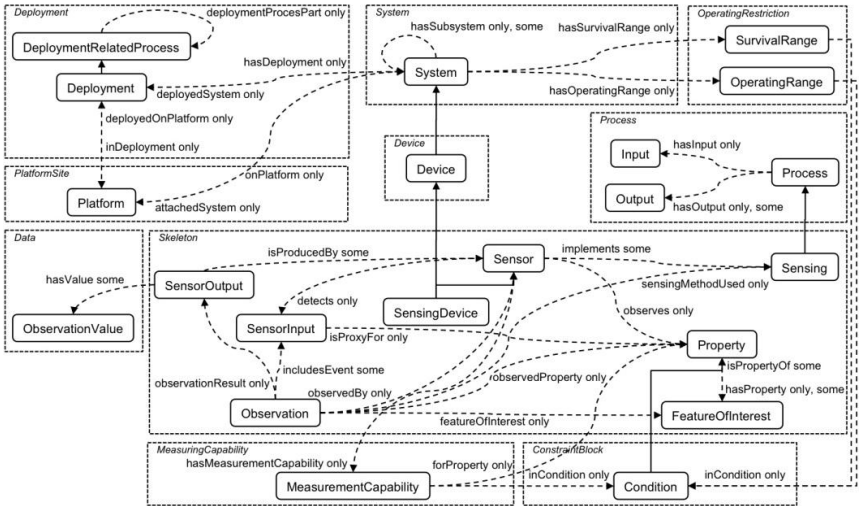

Fig. 1. SSN ontology for Sensor Network

A more precise, but still generic structure is necessary to facilitate the development of applications managing heterogeneous sensor data, like environmental or urban monitoring. A more precise structure simplifies the querying of data (easiness of query expression and query optimization). A generic structure allows a homogeneous management of data obtained from heterogeneous sensing systems.

A lot of existing works focused on physical sensor and real-time systems [9, 10, 11]. In our project, we rather focus on issues with heterogeneous data sources [6]. Query optimization and data indexing of sensor data are also issues in this context. In [16], the authors proposed a data model based on a proposition of spatiotemporal indexation of sensor data considering the most recent data. The model was also later used in [4, 5] and improved to support a methodology for evaluation of the quality of sensor data for the environmental phenomena monitoring systems.

These ontologies and models inspired our proposed data-centric VGS (Virtual Generic Sensor) model detailed in Section 4.1.

Concerning data management, the use of data warehouses for sensor data storage and analysis can be seen in various works like monitoring of pollinators [17], building energy and maintenance [18, 19], soil ecosystem [20]. The difficulties are that they only deal with physical sensor data and not with other heterogeneous data.

4. Results

In this paper, we present three contributions that support the methodology we describe in Section 2. Our contributions are: a generic model to represent heterogeneous sensor-like data sources and produced data, namely the VGS model; a declarative language to express simple and complex aggregated indicators on sensor data series; and a Web user interface to interactively explore those data.

4.1. VGS

Our first result is the VGS (Virtual Generic Sensor) model. Figure 2 shows the UML class diagram of this model. It describes the static structure of a generic acquisition system, with a **Sensor** composed of several **Detectors** (further detailed by **MeasureAttributes**, via the **MeasureType**), and a dynamic structure with **Samples**, produced by a Sensor, composed of **Measures** (further detailed as **Values**). **Deployments** correspond to campaign of measures, that take place to validate **Hypothesis**.

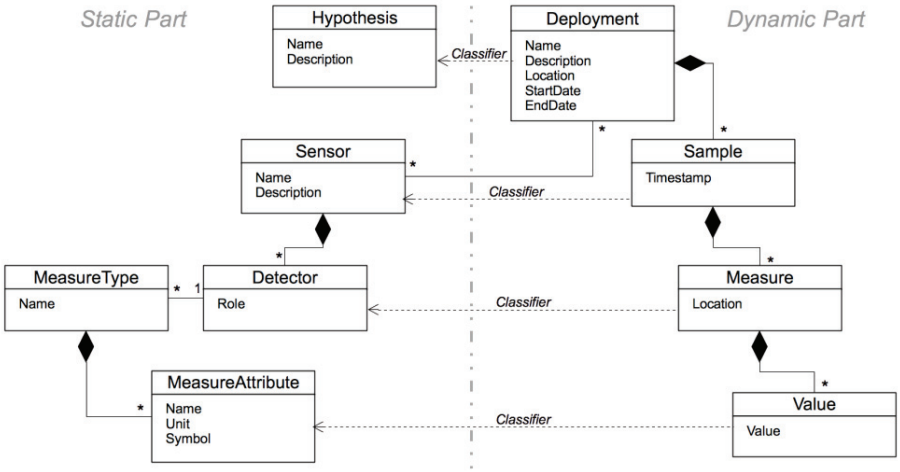

Fig. 2. VGS Model with UML representation

This model has been implemented for two Smart Building experimentation platforms managed by our project team: SoCQ4Home (illustrated with Figure 3), since October 2012; and MARBRE, since February 2014.

A total of around 400 heterogeneous physical sensors have been deployed to measure: temperature, humidity, CO2/VOC, contact (for doors/windows), electricity consumption, weather conditions... Physical sensor devices are modeled as VGS sensors, with detectors depending on the actual sensor type. A survey concerning fifty occupants of one of the buildings has also been realized. Results of questionnaires are going to be integrated: questionnaires are modeled as VGS sensors, and answers to questions are modeled as measures.

The current implementation of the VGS model is based on a MySQL database in particular to benefit from the expressiveness of the “golden standard” SQL language; moreover MySQL is a free open source tool.

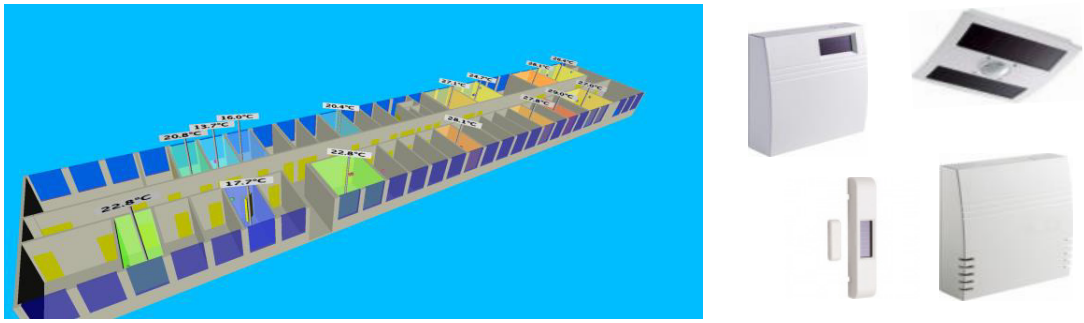

Fig. 3. SoCQ4Home Platform: Monitoring Building with Sensors

One of the major challenges involved in multidimensional data analysis [15] is to identify and define the dimensions and dimension levels of analysis. Dimensions are built with logical hierarchies. For example, Figure 4 shows a time dimension with one hierarchy. Three default dimensions are attached to the core VGS model: Time, with the timestamp attribute of Samples; Location, with the location attribute of measures; and Source, that represents the hierarchy of VGS concepts (Detector, Sensor, Deployment). User-defined additional dimensions can be added to further describe measures from generic sensors.

These dimensions are then used to specify data aggregation and data visualization levels. For example, Figure 4 shows the definition by a user of an aggregation level of data (*hour level* into the Time dimension) and a visualization level of data (*month level* into the Time dimension).

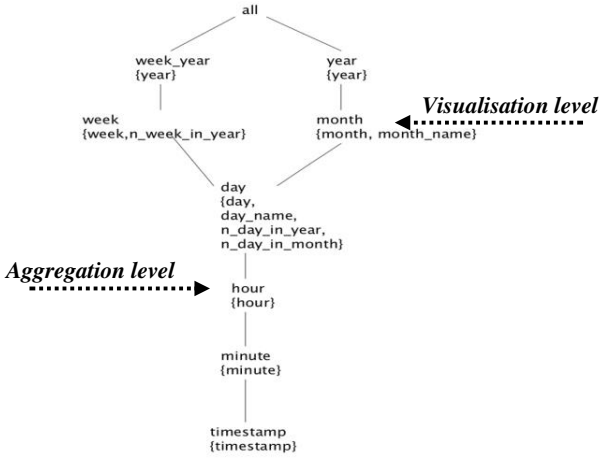

Fig. 4. Dimension hierarchy: difference between aggregation level and visualization level

4.2. Declarative language for indicators definition

Our second result is a formal model and a declarative language to finely define indicators as aggregations along dimensions for VGS data. It is based on the relational algebra (a foundation concept for relational databases, like SQL databases). In our current prototype, we implemented this language by automatically translating it to complex nested SQL aggregation queries. It enables to easily integrate new domain-specific dimensions and/or adapted existing dimensions (like Time and Location).

Due to space limitations, we do not describe the formal definition of this language. We sketch its expressiveness with an example: a user can define an indicator as a 2-step aggregation along the Time dimension for temperature sensor data, with an average at the minute level, and then a formula like “(MAX+MIN)/2” at the hour level, applied

to the previous average at the minute level. An indicator definition can also span over multiple dimensions, like Time and Source.

4.3. Visualization with a Web Interface

A web dashboard for the sensor data has been developed to visualize the various measures and indicators that describe a phenomenon. Figure 5 shows the SoCQ4Home dashboard of the administrator. It shows the temperature recorded during the last 48 hours and average temperature recorded during last 30 days, in user's office. The dashboard also shows the current temperature in some other representative rooms of the building. A visualization of building in 3D permits to project the results of exploration queries over the real geography of the building (see Figure 3).

Fig. 5. Dashboard for Smart Building

Our third result is a proof-of-concept Web user interface to visualize data and/or indicators as a matrix of graphs, according to the aggregation level of indicators (e.g., hour level in the Time dimension) and to the navigation level interactively defined by the user (e.g., month level in the Time dimension). The style of visualization is illustrated in Figure 6: one line per sensor in a graph (with one data point for each hour), and one graph per room (matrix rows) for Temperature and Humidity indicators (matrix columns). In order to facilitate visual correlations, graph scales are identical within a column, as well as the time axis between the 2 indicators. In this example, we can visually compare Temperature evolution between rooms, and visually search for potential correlations between Temperature and Humidity in each room.

The development of a full exploration Web interface, including definition of dimensions, indicators, and the dynamic navigation along dimensions is a work in progress.

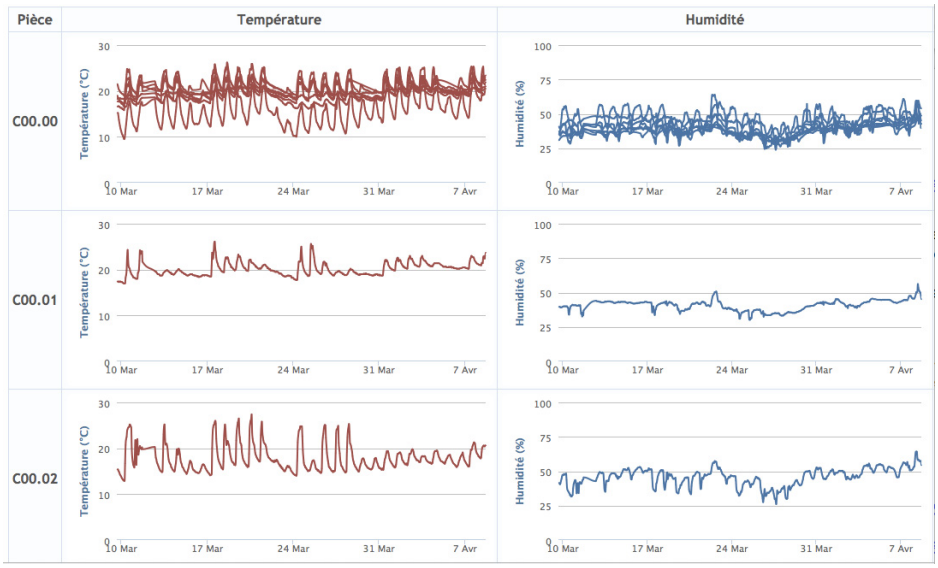

Fig. 6. Web user interface to visualize data as a matrix of graphs (actual data from MARBRE platform)

5. Conclusion

Our approach contributes to a better understanding of urban-related phenomena through cross-disciplinary analyses of large amount of data coming from phenomenon observations issued from multiple sources (sensors, surveys, various studies).

This approach aims at being well integrated with user-specific and domain-specific analyses processes, in particular for scientific data analyses. Analyses may be multi-dimensional through the definition of dimensions: spatial dimensions, temporal dimensions, and field-specific dimensions. Dimensions dedicated to a given phenomenon have to be identified and co-built by computer scientists and scientists from other urban-related disciplines. Our methodology is agile, incremental, iterative, and interactive to allow knowledge discovery along the way by users.

Our VGS model is generic as it enables heterogeneous data handling. The VGS model is semantically compatible with standard sensor ontologies defined by standardization organizations like Open Geospatial Consortium standards [1, 2, 3].

Our conceptual VGS model has been built from a real multi-disciplinary approach and its generic design makes it easy to apply to other phenomena observation. This model, as well as our agile exploration approach, is moreover independent from a specific data management technology. Although it is currently implemented on a SQL database, we aim at implementing it also on Big-Data-oriented databases like MongoDB or Cassandra.

Acknowledgment

This work was supported by the LABEX IMU (ANR-10-LABX-0088) of University of Lyon, within the program “Investissements d’Avenir” (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

[1] C. Reed, M. Botts M., J. Davidson, G. Percivall, Ogc® sensor web enablement: overview and high level architecture, Autotestcon, IEEE, 2007, pp. 372–380

[2] DUL: The DOLCE+DnS Ultralite ontology, 2010, HYPERLINK "http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite"http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite

[3] M. Compton M. & al., The SSN Ontology of the W3C Semantic Sensor Network Incubator Group, 2011, 7p. HYPERLINK "http://www.ict.csiro.au/staff/kerry.taylor/SSN-XG_SensorOntology.pdf"http://www.ict.csiro.au/staff/kerry.taylor/SSN-XG_SensorOntology.pdf

[4] CCG. Rodriguez, S. Servigne, Managing Sensor Data Uncertainty: A Data Quality Approach, in: International Journal of Agricultural and Environmental Information Systems, 2013, pp. 35-54
[5] C. Gutierrez Rodriguez, S. Servigne, R. Laurini, Towards real time metadata for networked based geographic database, in: Proccedings of ISSDQ2007, 5th International Symposium of Data Quality, Enschelde, 2007, 8p.
[6] G. Noel, S. Servigne, Indexation multidimensionnelle de bases de données capteur temps-réel et spatio-temporelles, in : Ingénierie des Systèmes d'Information. Vol.10, n°4, 2005, pp.59-88
[7] Y. Gripay, F. Laforest, J-M. Petit, A Simple (yet Powerful) Algebra for Pervasive Environments, EDBT 2010, 13th International Conference on Extending Database Technology, Lausanne, Switzerland, 2010, pp. 1-12.
[8] N. Lumineau, F. Laforest, Y. Gripay, J-M. Petit, Extending Conceptual Data Model for Dynamic Environment, In 31st International Conference on Conceptual Modeling (ER 2012), Florence, Italy, 2012
[9] P. Bonnet, J. Gehrke, P. Seshadri, Towards Sensor Database Systems, in Proceedings of Mobile Data Management, Conference, LNCS Springer, 2001, pp.3-14
[10] O. Diallo O. & al., Real-time data management on wireless sensor networks: A survey, in: Journal of Network and Computer Applications, vol. 35, n°3, 2012, pp. 1013-1021
[11] M. Kassim & al., A Based Temperature Monitoring System, in: International Journal of Multidisciplinary Sciences and Engineering, Vol. 2, 2011, pp.17-25.
[12] N.S.Patil & al., Data aggregation in wireless sensor network. International Journal of Service Computing and Computational Intelligence, vol. 1, no 1, 2011, p. 7-10.
[13] SensorML, OpenGIS® Sensor Model Language (SensorML) Implementation Specification. Open Geospatial Consortium, 2007, pp. 1–87
http://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/.SensorML
[14] SoCQ4Home, 2012, SoCQ4Home Project, HYPERLINK "http://liris.cnrs.fr/socq4home/"http://liris.cnrs.fr/socq4home/
[15] R. Kimball, M. Ross, The data warehouse toolkit: the complete guide to dimensional modeling, John Wiley & Sons, 2011
[16] G. Noel, S. Servigne, R Laurini, Spatial and temporal information structuring for natural risk monitoring, In: GISPlanet'2005, International Conference and Exhibition on Geographic Information, 2010, pp. 1-10
[17] R. A. G da Costa, tC. E. Cugnasca , Use of data warehouse to manage data from wireless sensors networks that monitor pollinators, in: Eleventh International Conference on Mobile Data Management, MDM 2010, Kanas City, Missouri, USA, 23-26 May 2010, pp. 402–406. IEEE Computer Society
[18] H-Y. Gökçe, Y. Wang, K. Gökçe, K. Menzel, A data-warehouse architecture supporting energy management of buildings, 2009, CIB W78.
[19] Stack, P., K. Menzel, D. Flynn, A. Ahmed, et S. McCormac, A cloud-based platform for energy and maintenance management, in: 4th International Conference on Computing in Civil and Building Engineering (14th ICCCBE), 2012, Volume 27, pp. 1-8.
[20] Szlavecz, K., A. Terzis, S. Ozer, R. Musaloiu-Elefteri, J. Cogan, S. Small, R. C. Burns, J. Gray, A. S. Szalay, Life under your feet : An end-to-end soil ecology sensor network, database, web server, and analysis service, 2007,CoRR abs/cs/0701170.