

Context-based Data Quality Metrics in Data Warehouse Systems

Flavia Serra, Adriana Marotta

Universidad de la República, Facultad de Ingeniería,
Montevideo, Uruguay, 11300
{fserra, amarotta}@fing.edu.uy

Abstract

The fact that Data Quality (DQ) depends on the context, in which data are produced, stored and used, is widely recognized in the research community. Data Warehouse Systems (DWS), whose main goal is to give support to decision making based on data, have had a huge growth in the last years, in research and industry. DQ in this kind of systems becomes essential. This work presents a proposal for identifying DQ problems in the domain of DWS, considering the different contexts that exist in each system component. This proposal may act as a first conceptual framework that guides the DQ-responsible in the management of DQ in DWS. The main contributions of this work are a thorough literature review about how contexts are used for evaluating DQ in DWS, and a proposal for assessing DQ in DWS through context-based DQ metrics.

Keywords: Data Quality, Data Warehousing System, Data Warehouse, Context.

1 Introduction

Data Quality (DQ) is a wide research area, which includes different aspects, issues and challenges. Besides, it is extremely relevant for industry due to its impact in information systems of all application domains. DQ refers to a set of characteristics that data should have, such as accuracy, currency, completeness, etc. [1]. The consequences of low quality in data are often experimented in daily life, for example when a wrong address or duplicate customer information generate problems with customer relationship. On the other hand, when considering multiple and heterogeneous sources that must be integrated, DQ of the sources is usually unknown and must be evaluated and improved as part of the integration process [2].

The fact that DQ depends on the context in which data are produced, stored and used, is widely recognized in the research community [3]. Following this idea, the *fitness for use* approach, firstly presented in [4], is also a widely adopted concept, where quality is considered as totally dependent on the adequacy of data to their intended use.

The *Context* research area is also very wide and has many different applications. Many authors have already identified the importance that context has for data, for example in [5]. For instance, in Pervasive Computing domain, where applications operate in highly dynamic environments, context is necessary for selecting data relevant to the user, as these data should be adequate to the user environment, and this environment determines the context [6].

On the other hand, the Data Warehouse research area had an enormous development during the last two decades, in particular the topics of design, loading and maintenance, and OLAP (On-Line Analytical Processing). At the same time, there was a significant grew in the use of these systems in industry and a proliferation of tools for their implementation and exploitation. Moreover, in the last years, the increasing volume of available data and the importance data have gained, have contributed to the relevance of the Data Warehouse Systems (DWS), whose main goal is to give support to decision making based on data. DQ in this kind of systems becomes essential. It can be addressed from two points of view: (i) the final user should be aware of the quality of the data he is receiving, and (ii) the system should be able to control and improve the quality of the data it receives, processes and provides.

This work presents a proposal for identifying DQ problems in the domain of DWS, considering the different contexts that exist in each system component. This proposal may act as a first conceptual framework that guides the DQ-responsible in the management of DQ in DWS.

The main contributions of this work are the following: (i) a literature review identifying and analyzing the proposals that relate the topics of DW, DQ and Contexts, (ii) the definition of contexts in DWS and DQ metrics based on these contexts, and (iii) a case study showing the application of the proposal.

This paper is an extension of [7], where we presented a reduced literature review and one metric for each kind of DQ defined. In this paper we present a wider literature review where we consider many more works and

additionally, we present some relevant results summarized into tables. On the other hand, the present work adds more DQ metrics that show different strategies for measuring DQ with context-based metrics.

The rest of the paper is organized as follows: Section 2 introduces the three research areas involved in this work, Section 3 presents a literature review, Section 4 presents an approach for assessing DQ taking contexts into account, Section 5 presents a Case Study where a set of DQ metrics are defined, and finally, Section 6 presents the conclusions and future work.

2 Data Quality, Data Warehouse Systems and Contexts

This section presents the main concepts of the research areas over which this work is built: Data Quality, Data Warehouse and Contexts.

The impact that low DQ has over the organizations and the huge influence it has over all kinds of information systems in all kinds of application domains is a widely recognized problem. However, the meaning of *data quality* is not always understood and there are still in the literature different definitions for it. In [8] DQ is defined as the capacity of satisfying the requirements for the use of data. According to the authors, DQ depends on the intended use of the data as well as on the data itself. In [9] DQ is defined as the *fitness for use* of the data. This concept is widely adopted in the DQ related bibliography, such as [1][3][10][11]. In [1] the authors present an analysis of the different definitions of DQ that were proposed since the 90's, and they remark that a definition including the main characteristics considered by most of the existing proposals is needed. They conclude that DQ refers to a set of quality dimensions, which in general are defined as quality properties or characteristics, and which are grouped in four categories: accuracy, currency, completeness and consistency. In addition, they state that these are the dimensions more often associated to DQ. However, they say that there is not an agreement in the literature about the set of dimensions that characterize DQ or the meanings of them, and a standard has not yet been established.

Among all the existing kinds of decision support systems, DWS are the ones over which academic and industrial communities had paid more attention [12]. In [13], Inmon gives the following definition: “A DW is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process”. The data offered by a DW are extracted from various sources, and then integrated, cleaned and transformed in order to be used at the moment of decision making [12][13]. One big challenge of this kind of system is to achieve the integration of multiple heterogeneous data sources.

DW are based on the multidimensional model [14]. This model allows a better understanding of data and leads to better performance in complex queries. In multidimensional models data are presented in an n-dimensional space, generally known as *data cube* or *hypercube*, where the axes are the dimensions of analysis and the points in the space contains the indicators, called measures. Each coordinate of the cube is called a fact. The dimensions are structured in hierarchies that give the criteria for data aggregation, which is an essential operation in multidimensional analysis. Besides, the hierarchies are composed by levels, and the instances of a level are called members. For example, in the dimension Time there may be a hierarchy composed by levels date, month, quarter, year. ‘Q1’, ‘Q2’, ‘Q3’, ‘Q4’ are members of the level quarter. Fig. 1 shows an example of a data cube [14].

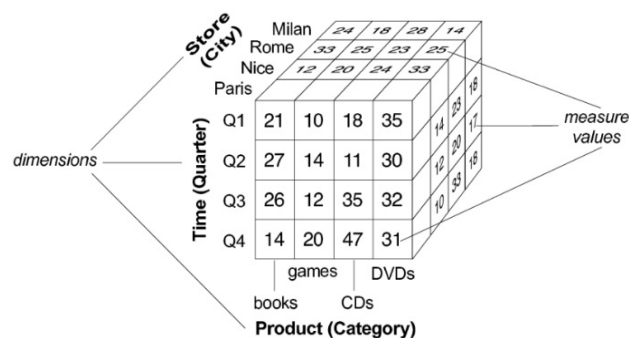


Figure 1: Three-dimensional cube of sales data, with the dimensions *Store*, *Time* and *Product*, and the measure *Amount*

A DWS is an information system whose main component is a DW, and whose general architecture is composed by different components that make possible the whole process from the data sources towards the end-user. In [12] different architectures are presented, but the two-tiered architecture, shown in Fig. 2, is stated as the

most referenced one. It remarks the separation between the sources and the DW, although it represents the four main stages in the data flow:

- *Sources*: They contain the raw material of a DWS and they may be internal to the organization (operational systems, e-documents, etc.) or external to it (web data, other organization's data, etc.).
- *Data staging*: ETL (Extraction, Transformation and Loading) processes are used for loading the data obtained from the sources into the DW. The data staging is a temporal storage for extracted data, over which all data integration and transformation processes are executed before data is loaded into the DW.
- *Data Warehouse*: The DW includes all corporate data, while a Data Mart (DM) contains a data subset or aggregated data from the DW and it is designed for a specific organization sector, i.e. for a specific analysis domain. The metadata contains information about the DW data, such as sources information, access procedures, users information, data provenance, etc. [12].
- *Analysis layer*: The OLAP server provides the needed support for multidimensional data organization and multidimensional data manipulation, allowing the execution of multidimensional operations over data [15]. There are a variety of possible client tools, which allow the users to do interactive analysis, such as OLAP, data mining, and specific analysis tasks.

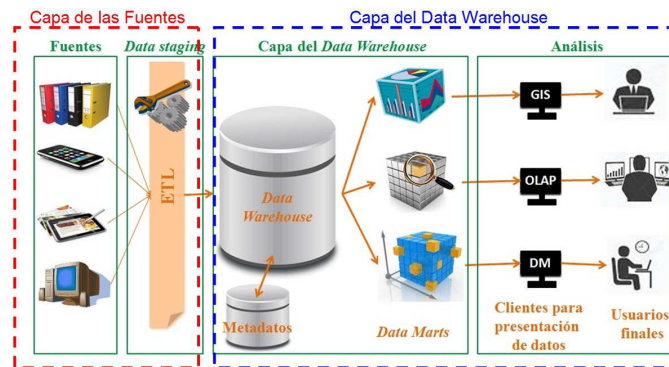


Figure 2: Two-tier architecture in Data Warehouse.

As said before, the importance of considering the context in DQ management is widely recognized. However, it is not possible to find in the literature a concise and globally accepted definition for the concept of context. In fact, there are a lot of conceptualizations and definitions for it, whose approaches depend on the research domain where it is applied. For example, in [16], the context is defined as the possibility of selecting data according to the user environment, while in [5] the authors consider that the context is a set of variables of interest that influence the actions of an agent. In [17] a set of definitions extracted from the Web are analyzed. The authors consider that is difficult to find a relevant definition that satisfies all disciplines and they mention that there are still few ideas about the relevant properties that should be considered when modeling context.

3 Literature Review

The goal of this literature review is to find scientific works where context is defined for DQ assessment in DWS. In order to achieve this, a main question is posed, which motivates this research:

How contexts may be used for assessing DQ in DW?

Since this question is too specific and restricts too much the search space, we consider the possibility of posing more general questions that involve a wider set of works, without losing the research focus. Those questions and their results are presented in the following.

3.1 Analysis of research questions

How DQ is managed in DWS?

DQ in DWS is a very much studied issue. Many works are focused on solving DQ problems in the stage of data sources identification and selection, and in the ETL stage. However, the existing work shows that it is not clear yet which is the best stage for addressing DQ assessment in this kind of systems.

The work presented in [18] focuses on measuring believability of data. It states that believability depends on the provenance of data, which refers to the sources of the DWS. In [19] the authors present DQ as a challenge in DW design and implementation. They remark the importance of quality and consistency checking in the data that comes from the sources. In this work data consistency is presented as a separate concept from DQ, differently from most works that present consistency as a particular aspect of DQ.

In [20] the authors address the quality process as part of the ETL tasks, presenting a proposal for DQ control and data cleaning. They propose rules for data transformation, which do the data cleaning, and a rule generator that facilitates the inclusion of DQ processes into the ETL tasks. In [21] they also focus on data cleaning in the ETL process, proposing a framework of configuration rules, which has two inputs: a data profiling used for selecting the data sources to be cleaned, and user opinions about already known data problems. In the same direction, the authors of [22] state that the ETL data cleaning tasks should identify the data problems and inconsistencies and perform the necessary transformations for assuring a minimum DQ for the data that will be loaded in the DW.

With respect to the data analysis stage, in [23] the authors claim that DQ is subjective from two perspectives: on one hand, DQ problems may be more or less relevant depending on the decisions to be made by the data user, and on the other hand, data analysts may have previous personal knowledge and opinions about the quality of the received data. In [24] the authors emphasize the influence of the quality of data management in business intelligence (BI) systems throughout the whole process from the sources to the decision making stage. In [25] the impact that DQ has on decision making, organizations believability and customer satisfaction is remarked. They consider that DQ problems in multidimensional repositories still need to be correctly organized and addressed.

In summary, the importance of DQ in DWS has been widely shown in the existing literature, since many researchers present the challenge of managing DQ in these systems. However, there is not consensus about how this task should be done. Most of the works have only addressed the problem of data cleaning in the ETL stage, ignoring the task of DQ assessment throughout the whole DWS lifecycle (or DWS components). In addition, although many works have made efforts on identifying DQ dimensions for DWs, a consensual set of DQ dimensions has not been identified yet, questioning the possibility of defining a specific and unique set of DQ dimensions for DW environments.

How Contexts are defined and used in DWS?

DWS are determined by the dimensions and measures that describe the business. Based on this, some authors consider the context as the dimensions that allow the analysis of the measures. In particular, in [19], the authors present the DW facts as the numerical data needed for satisfying all calculation options that are of interest to the end user, and they say that dimensional tables provide context to these facts. Considering this concept, the stage of requirements elicitation and conceptual analysis in the DWS design, has great influence on the context that will be given to the DW measures by the DW dimensions.

Some works focus on the stage of source identification and selection [26][27][28][29], stating that organizational data coming from traditional DW sources, i.e. operational data, can be more successfully exploited if they are combined with data external to the DW. In [28][29] the authors consider that the context is defined by documents content. According to [28], the most common information sources for users are the intranet of their organization, the web, and the emails. These unstructured data are often related with the DW entities and relationships, constituting the context.

Meanwhile, in [30] the authors consider the context in the ETL stage, grouping the users that share interests in this stage and contextualizing the ETL process according to the different points of views of the experts.

In the data analysis stage the use of contexts is also emphasized. In [29] they claim that the context that allows the understanding of the data stored in the DW is usually present in separate documents, whose value is not correctly exploited. These documents provide information about the DW facts, describing their context. In [31], with another approach for the concept of context, the authors propose the use of rules for representing knowledge as a mean for taking context into account during OLAP. In [28][29] the authors state that the OLAP tasks determine the context of the decision making process, and this context is composed by all the information involved by these tasks. They claim that contextual information must be taken into account during DW exploitation. However, although several context-sensitive applications have been proposed, there is not enough research in the integration of context to DWS yet.

Finally, we also studied other works from which we obtained relevant results, but we do not detail here [32][33][34][35][36][37][38][39][40][41][42][43].

How Contexts are considered in DQ assessment?

The concept of *fitness for use* is widely used in DQ literature. Several authors consider that this concept means that DQ depends on the context where data are used. The works presented in [21][44][45][46][47][48] are based on or remark this approach. In [48] they explain that the information-consumers' needs and the circumstances around the use of the information should be taken into account. In [46] the authors mention that most of the existing approaches depend on the context, however the contextual dimension is barely represented in DQ frameworks.

The work in [49] shows that besides users may have different data requirements, their satisfactory quality level and the elements that define their context vary according to each-one's perspective. In this case, the context is determined by the users' requirements. Also in [50] the authors focus on users and their needs. In this case, the contextualized object are the queries made by users, and the context for this object is determined for users information (geo-location, habits, interests and needs), and queries data and processing quality.

In [51] the authors explore the contextual nature of the information and they use the term *Context* referring to a group of users with a specific purpose in a specific task. They consider that information quality should be evaluated taking the information construction process into account, not using a static measure. They propose to consider the states from data to knowledge, going through the different contexts. Additionally, in [48] DQ and information quality are considered different. In particular, for the authors, information quality metrics are necessarily context-dependent while DQ metrics may be absolute.

The work in [47], like many works found in the literature, is based on the DQ dimensions classification given in [3], where DQ dimensions are classified into: intrinsic (based on the degree to which data values adjust to real values), contextual (based on the degree to which data are applicable to user's task), representational (based on the degree to which data are presented in an intelligible and clear way) and accessibility (based on the degree to which data are available). For example, accuracy and objectivity are objective dimensions since they are intrinsic to the data and independent from the context where they are used. However, not all DQ dimensions can be objectively measured, since dimensions such as relevance and believability tend to vary according to the context where they are used. In [47] the authors claim that due to the dependence of DQ on the context, despite the wide discussions about DQ dimensions existing in the literature, it does not exist a unified set of DQ dimensions.

Finally, we also studied other works from which we obtained relevant results, but we do not detail here [52][53][54][55][56][57].

How contexts may be used for assessing DQ in DW?

Concerning the main question presented at the beginning of this section, only a few works are found, which intend to answer it.

In [45] a proposal where DQ is integrated to the whole DW development process is presented. They mention that a DQ dimension may involve objective information about data and their transformation process, but they also may be subjective, since different users may have different quality requirements. They also refer to the concept of *fitness for use* for describing DQ, highlighting the relateness of quality and the fact that certain data may be suitable for certain use and not for another use. This work shows the importance of considering context, although they do not use this term, when evaluating DQ in DW.

In [44] the authors present a proposal for managing the data cleaning process in the ETL stage. They define DQ in DW as the degree to which data satisfy the specific user needs, and as users' interests vary, they state that DQ needs to be adaptable. They implicitly consider user context for evaluating DQ. For the authors of [58], DQ is context-sensitive by nature and must be evaluated in the context of the business domain where data will be used. Additionally, they consider that DQ evaluation research has still much to do, as a relevant set of DQ factors for decision support systems has not yet been identified and the context has not been considered for DQ measurement in these systems. In these two works, the authors mention that additional factors should be considered in decision support systems, such as the role of the decision makers, their preferences and interests, and the hierarchies of the decisions. The authors remark the importance of user and decision contexts for evaluating DQ.

Finally, we also studied other works from which we obtained relevant results, but we do not detail here [59][60][61][62].

3.2 Relevant Results

In this section we present other results that are considered of interest to the general problem of the evaluation of DQ taking into account the context in DWS.

In Table 1 we show a classification of articles according to the definition of context that different authors present in the literature. The table shows for whom or what the context is defined and how this context is determined. Hence, the column *Object (contextualized object)* corresponds to the concept of *for whom* or *for what* the context is defined. In the column *Context (contextualizing object)* we show what defines the context and the column *Context Components* describes how such context (or contextualizing object) is composed.

Table 1: Contexts

Object	Context	Context Components	Articles
Aggregation	Data cube (user view on multidimensional data), DW measure and DW dimensions	Data cube information, user preferences and aggregation function information.	[31]
	Data	Data's dimensional aspects, relations between hierarchical categories, dimensional rules and constraints.	[59]
	Environment	Information of the environment, the user and the task that will be performed.	[46]
Data or Information Quality	Decision making	Decision categories (i.e. credit, tactical, etc.), weight of DQ factor over the decision category and the user confidence for each DQ factor.	[58]
	Users	User requirements.	[52]
	Users	The purpose of users in a specific task.	[47][51][53][63]
Queries	User and specificities Request	User information (geo-location, habits, interests and needs). Similar queries. Data and processing quality.	[50]
	Users	Features and user preferences.	[32]
	Environment	Source identification (sensor), identification of the entity for which the sensor picks up data, time in which data is collected and data.	[54]
Data	Data quality, Data and Sensor	Features (CPU, memory, energy, etc.) and sensor measures, user preferences, task information to be performed.	[55]
	Quality Requirements	Constraints defined by rules.	[56]
Multidim. Recommendation System	Users	User profile and his preferences	[60]
Data Sources	Sources structure	All data sources with similar structure (columns, records, etc.).	[57]
	Documents	Documents content.	[29]
DW measure	DW Dimensions	Dimension data.	[19][27][33][34][35][36][37][38][39][61]
ETL process	Users	Items shared among experts with equal interests in different stages of the ETL process.	[30]
Task	User	User information.	[40]
	Task	OLAP task information: document content, document metadata (author, creation date, etc.), preferences and user profile.	[41]
Decision Making	Environment	Kind of decisions and information needs of the user.	[62]
	Users	User preferences and specifications (DW facts and measures that have been selected).	[42]
	Data quality	Aspects of data/information (completeness, accuracy, relevance, etc.)	[48]
User	Documents	Document content.	[43]

We can see, In Table I, that there is a diversity of components for defining the contextualizing object. This is because the contextualizing object is not unique and it depends on the contextualized object. For example, in [52] the contextualized object is data quality, the contextualizing object is the user and it is determined by the user requirements. On the other hand, in [50] the authors consider that the large amount of information generated by users and applications is often underused. Meanwhile, the information retrieved from the data sources is often irrelevant to the user needs. For this reason, the processing of complex requests on such data sources is costly and does not guarantee the user satisfaction. For all this, the authors propose to take into account the context of the queries submitted to the different data sources.

The works presented in [54][55] are focused on *sensed contexts*, concept that is used in ubiquitous (or pervasive) computing environments. In this kind of environment, where data sources are sensors, taking into account the context is of great importance. Also in [57] the authors focus on an environment of multiple sources, and they consider that the same entity is stored redundantly in different data sources. Therefore, authors consider that entities with the same structure represent the same entity, and such structure (or syntax, according to the authors), determines the context of data sources. On the other hand, in [60] the contextual information is integrated into a multidimensional recommendation environment, to be used in OLAP systems.

Finally, an interesting approach, shared by various authors, is the one presented in the works that define the context of the DW measures as the corresponding DW dimensions [19][27][33][34][35][36][37][38][39][61]. DW measures are data and therefore these articles could be grouped with articles that consider data contexts. However, it is important to consider these works separately because they present a specific approach. Moreover, in [29] authors also focus on the DW measures context. However, in this case, the context is determined by the relevant documents of the organization and its contents.

We extract from Table I the most important contextualized objects for our work. In Table 2 these objects and the quantity of analyzed articles that focus on them are listed. On the other hand, in Table 3, it is interesting to spotlight the used contextualizing objects. In this case, we also show the most important objects for this research.

Table 2: Contextualized objects

Contextualized object	Total
Data/information quality	8
Data	3
Data source	1
DW measure	11
ETL process	1
Task	1
Decision making	4
User	1
Others	4

It is interesting to note how objects that are contextualized in some cases, in other cases are contextualizing objects. An example of this is observed for the following objects: *Users*, *Quality of data/information* and *Data*. We can see that the majority of the articles analyzed intend to give context to the DW measures and the Data/Information quality, while the DW dimensions and Users are most frequently used by researchers to give context to different objects.

Table 3: Contextualizing objects

Contextualizing object	Total
Environment	3
Data/information quality	2
Data	3
DW dimensions	11
Documents	2
Quality requirements	1
Task	1

Contextualizing object	Total
User	6
Others	4

4 DQ Assessment in DWS based on Contexts

The conclusions posed by the authors of the analyzed bibliography are our starting point to develop a proposal that defines a framework where data quality problems can be identified in Data Warehouse Systems. Besides, it is interesting to consider this framework as the basis for quality management. For this, we take into account the different components that are part of this kind of systems (shown in Fig. 2) and the different contexts that have influence over them.

One of the purposes of this work is to define contexts, present throughout all lifecycle of a DWS. These contexts are of interest for the data quality assessment, since they may affect it. Therefore, this proposal is focused on defining the different contexts that data go through, from the moment they are loaded in the DW until they are used by the end users. The view that data go through different contexts while being transported from the sources to the users is shared with the authors of [51]. On the other hand, it is important to indicate that this work does not focus on data sources and ETL layers quality issues, since many researches have already addressed them. For example, the works in [19][21][27] focus on data quality in DW sources, and the works in [21][44][63] focus on quality, especially data cleaning, in ETL processes. In particular, the works in [30][64] focus on the contexts of the ETL processes and the data sources, respectively.

4.1 Context in DWS Components

In this section, we present and define the contexts for the components of the DW layer (shown in Fig. 2). Each component, with its respective context, is presented in Fig. 3. In the following we describe the elements that determine each one of these contexts.

- *Context in the Data Warehouse (DWC)*: It is defined by data in the DW, and also documents, e-mails and other data external to the DW that are related to data stored in it. The external data may be tables of personal databases of different users within the organization, which may be linked to the DW data. On the other hand, documents and e-mails that are shared within the organization also may contain information strongly related to data stored in the DW.
- *Context in the Data Mart (DMC)*: A Data Mart contains a subset of the data stored in the DW, which have been transformed, and is directed to a specific analysis domain (e.g. a section in the organization). Hence, for us, the DMC is determined by a set of rules that describe properties, constraints and quality requirements specific to the corresponding analysis domain.
- *Context in Use (CiU)*: The CiU is the context in the data presentation component, and is determined by data that describe the end-user. These data may be geographical location, language, role, requirements (of data or quality), etc. For example, the DQ requirements could be a minimum level of data accuracy or data completeness.

Information regarding contexts could be stored in the metadata repository shown in Fig. 3. An example of such information are the domain rules, which determine the context in the Data Mart component.

4.2 Data Quality according to Contexts

For the quality assessment in the DWS components, taking into account the contexts presented above, this work is based on two quality approaches that are presented below. Fig. 4 shows where these approaches are applied in the DWS.

- *Crosby's Meeting Requirements* [65]: This approach emphasizes the compliance with the system requirements and it focuses on prevention rather than correction, where the only way to achieve good performance is having zero defects. The essence of this approach is that the requirements must be known and they must be translated into measurable characteristics of the product/service. Therefore, the quality of a product/service is equivalent to the satisfaction of the specification criteria by all the measurable characteristics of such product/service [66]. The Meeting Requirements approach is applied to evaluate data quality in the Data Warehouse and Data Marts components.

- *Juran's Fitness for Use* [4]: This approach is focused on meeting the needs of the user through the suitability (of the product) for the intended use. The approach called Fitness for Use is applied to assess data quality in the data presentation layer.

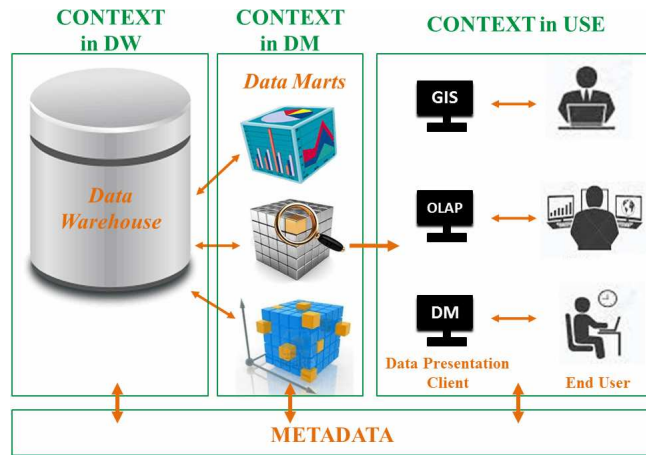


Figure 3: Contexts in DWS.

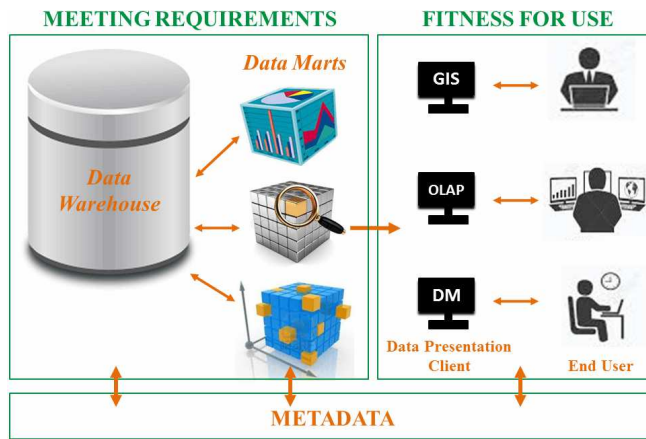


Figure 4: Two quality approaches in DWS to evaluate Data Quality according to the Context.

Therefore, from the quality approaches presented above, we define quality in components of the DWS, as shown in Fig. 5.

- *Quality in the Data Warehouse (DWQ)*: It depends on the DWC, therefore the data quality metrics for the Data Warehouse component are defined using this context.
- *Quality in the Data Mart (DMQ)*: It depends on the DMC, therefore the data quality metrics for the Data Mart component are defined using this context.
- *Quality in Use (QiU)*: It depends on the CiU, therefore the data quality metrics for the data presentation component are defined using this context.

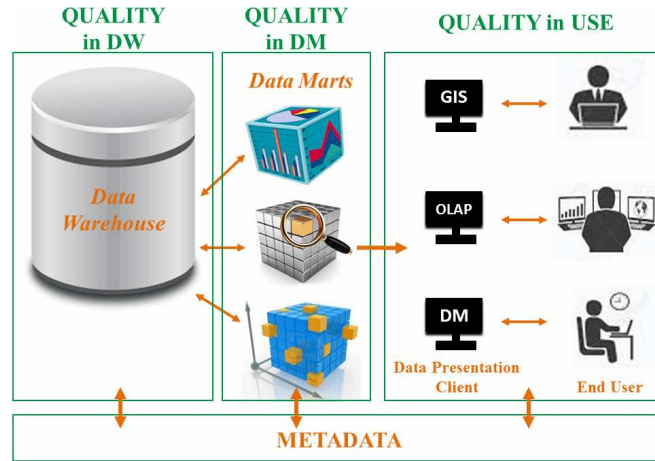


Figure 5: Quality in DWS.

The concept of quality applied in the data presentation layer of the DWS coincides with the definition of quality in use of ISO/IEC 25010 [67]. This norm defines *quality in use* as the quality from the point of view of the user. In addition, the standard defines a set of quality features, some of which are divided into sub-characteristics. An example of the quality feature is usability and the set of sub-characteristics for it, according to the standard, are effectiveness, efficiency and satisfaction. However, these are not the most interesting concepts for this research, but the most relevant is that the *quality in use* model assesses the quality in a particular context of use that depends on the point of view of the user.

Although the ISO/IEC 25010 [67] focuses on a quality model for Software, it can be adapted and applied to other areas. An example of this is the application of this norm for the definition of a quality in use model in Web portals [68].

5 DQ Metrics Proposed for a Case Study

In this section, we provide a case study that allows applying the concepts introduced before for defining the DQ metrics. The example used represents a supermarket chain that holds information about sales and promotions in a DW. There are two domains of data analysis, *Sales* and *Advertising*, in which different requirements are manifested. These requirements originate the conceptual multidimensional design described in the following paragraph.

In Fig. 6 and 7 the conceptual schema corresponding to the example is presented. We use the CMDM model [69] to represent the dimensions and the dimensional relationships. Firstly, Fig. 6 shows the dimensions involved in the case study: *Product*, *Time*, *Branch* and *Promotion* with their respective hierarchies. In CMDM each dimension is represented as a box that includes hierarchies, which are composed by levels connected by arrows that represent a N to 1 relationship. In Fig. 7 the dimensional relationships *Sales* and *Discounts* are presented. In CMDM dimensional relationships represent the crossing between dimensions and the involved measures (identified by an arrow) that originate the data cubes, i.e. a dimensional relationship represents all the data cubes that may be built, whose axes are formed by one level of each dimension represented in the relationship.

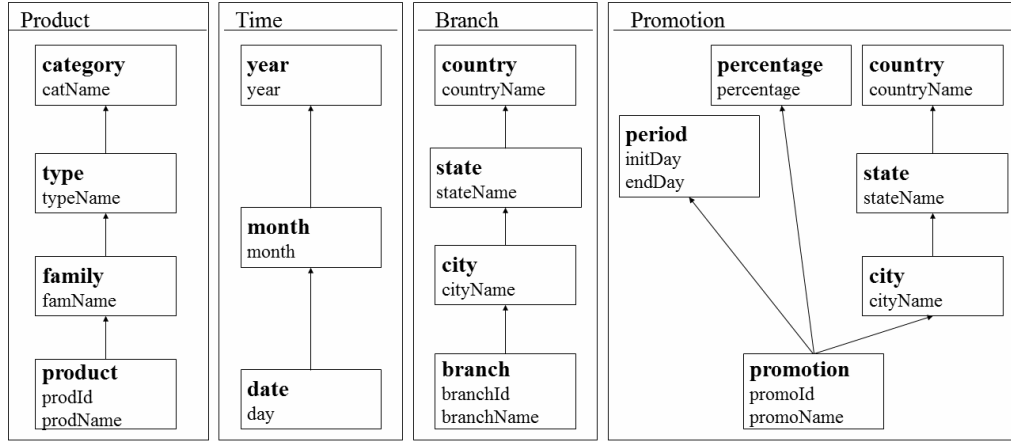


Figure 6: Hierarchies with their levels for each dimension.

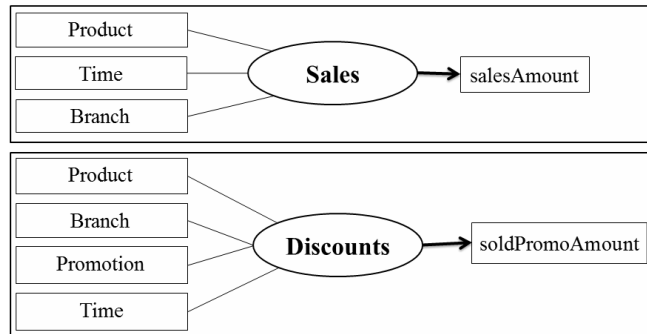


Figure 7: Dimensional relationships *Sales* and *Discounts*.

Considering the DW implementation in a relational database, Figs. 8, 9, 10 and 11 show examples of instances of the dimension tables *Product*, *Time*, *Branch* and *Promotion*, respectively. In each tuple of the dimension table *Product* (Fig. 8), the product identifier, its name, the family to which it belongs, the product type and its category are presented. In the dimension table *Time* (Fig. 9), each tuple contains the date, the month and the corresponding year. In the dimension table *Branch* (Fig. 10), each tuple contains the identifier of the branch, its name, its city, state and country. Finally, we have the dimension table *Promotion* (Fig. 11), where each tuple contains the promotion identifier, its name, its city, state and country, the duration period of the promotion (starting and end dates) and the discount promotion percentage.

prodId	prodName	famName	typeName	catName
23	chocolateBar	chocolate	food	premium
15	chocolatesIceCream	iceCream	food	premium
40	chocolateTruffles	chocolate	food	premium
...	...	...	...	...
56	toffee	candies	food	premium

Figure 8: Dimension table *Product*.

day	month	year
03/01/2014	Ene-14	2014
...	...	...
05/04/2015	Abr-15	2015
06/04/2015	Abr-15	2015
07/04/2015	Abr-15	2015
08/04/2015	Abr-15	2015
...	...	...
10/11/2015	Nov-15	2015

Figure 9: Dimension table *Time*.

branchId	branchName	cityName	stateName	countryName
25	Supermarket	Maldonado	Maldonado	Uruguay
26	Super_Salto_26	Salto	Salto	Uruguay
28	Super_Florida_28	Florida	Florida	Uruguay
29	Florida_29	Florida	Florida	Uruguay
30	Florida_30	Florida	Florida	Uruguay
...	...	...	...	...
31	Super_Rocha_31	Rocha	Rocha	Uruguay

Figure 10: Dimension table *Branch*.

promold	promoName	promoCity	promoState	promoCountry	initDay	endDay	percentage
p1	Children's Day	Florida	Florida	Uruguay	01/01/2014	03/01/2014	20%
p2	Father's Day	Colonia	Colonia	Uruguay	10/04/2015	15/04/2015	50%
p3	Mother's Day	Maldonado	Maldonado	Uruguay	09/08/2015	11/08/2015	10%
...	...	...	...	...	...	...	...
p56	End of season	Salto	Salto	Uruguay	09/12/2015	12/12/2015	18,03%

Figure 11: Dimension table *Promotion*.

On the other hand, the fact tables *Sales* and *Discounts* are shown in Fig. 12 and Fig. 13, respectively. The former has an identifier, the product identifier, the sale day, branch identifier and the sold amount of the product. The latter has an identifier, the promotion identifier, the product identifier, the sale day, branch identifier and the sold amount of the promotion.

saleId	prodId	day	branchId	saleAmount
v1	23	03/01/2014	26	50
v2	15	03/01/2014	26	20
v3	40	03/01/2014	28	35
...	...	...	...	...
v50	56	05/04/2015	28	35

Figure 12: Fact table *Sales*.

disId	promold	prodId	day	branchId	soldPromoAmount
d1	p1	15	03/01/2014	28	35
d2	p2	56	29/03/2015	31	5
d3	p3	40	05/08/2015	29	20
...	...	...	...	...	...
d40	p56	23	10/10/2015	26	10

Figure 13: Fact table *Discounts*.

5.1 DQ in the DW

In this section we show how the different elements that give context to the DW component determine the quality in this component. In the following, two examples are presented.

5.1.1 Example 1. Context: Document of the organization

In this example, the context is determined by documents that belong to the organization. Table 4 shows, for the quality dimension *accuracy* and its quality factor *semantic accuracy*, the quality assessment of each tuple regarding the attribute *city*, in the dimension table *Branch*. This attribute represents the city where each supermarket's branch belongs. A document of the organization acts as a referential for the assessment. It contains for each city the corresponding list of branches that belong to it. Part of the document corresponding to the supermarket *Super* is shown in Fig. 14.

Table 4: DQ Metric *dwq_Example1*.

<i>Component</i>	DW
<i>DQ Dimension</i>	<i>Accuracy</i>
<i>DQ Factor</i>	<i>Semantic accuracy</i> : It refers to accuracy of the data regarding the real word.
<i>DQ Metric</i>	<u>Name: <i>dwq_Example1</i></u> <u>Contextualizing object</u>: Organization's document, which contains for each city the corresponding list of branches that belongs to the same. <u>Contextualized object</u>: Dimension table <i>Branch</i> <u>Granularity</u>: Tuple <u>Description</u>: Check each tuple in the dimension table <i>Branch</i>. For each branch <i>bi</i> of the table, it considers its city <i>cx</i> in the tuple. Later, it searches the city <i>cy</i> in the referential document, such that this is the city to which <i>bi</i> belongs effectively. Finally, if <i>cx</i> = <i>cy</i> then the tuple is correct. Otherwise, it is an incorrect tuple for the branch <i>bi</i>. <u>Result type</u>: {0,1}

• ...
• Florida city
– Id: 28, name: Super_Florida_28
– Id: 29, name: Florida_29
– Id: 30, name: Florida_30
• Rocha city
– Id: 31, name: Super_Rocha_31
– Id: 32, name: Super_Rocha_32
• Salto city
– Id: 25, name: Supermarket
– Id: 26, name: Super_Salto_26
– Id: 27, name: Super_Salto_27
• ...

Figure 14: Structure of the document that contains the branches in each city.

For the quality assessment of the dimension table *Branch* with regard to the attribute *cityName*, we verify that each city stored in the dimension table, is the city where the considered branch effectively belongs. For example, in Fig. 10 it can be seen that:

- *branch* 'Super_Rocha_31': this branch belongs to the city 'Rocha', according to the referential document of the organization. The column *cityName* for this branch says 'Rocha', hence this tuple is correct.

- *branch ‘Supermarket’*: according to the referential document of the organization, this branch belongs to the city ‘Salto’. However, the column *cityName* for this branch, in the dimension table *Branch* says ‘Maldonado’, hence this tuple is incorrect.

Once all tuples in the dimension table *Branch* have been evaluated, the values obtained from the DQ metric *dwq_Example1* are saved in the metadata repository. These metadata can later be used. For example, one of their utilities could be the correction of the dimension table *Branch*, to have the correct city to which belongs each branch.

5.1.2 Example 2. Context: DW dimensions

In this example, the context is determined by DW dimensions. In Fig. 15 we present the dimension tables that determine the context in this case. From dimension table *Branch* we consider the level *cityName* (this contains the cities where the branches belong) and from dimension table *Promotion* we consider the level *cityName* (this contains the cities where the promotions are performed).

Table 5 shows, for the quality dimension *consistency* and the quality factor *intra-relation integrity*, the quality assessment of the fact table *Discounts* regarding the city of the branch in which the promotion was performed and the city where the promotion was performed effectively. In this case, the DQ metric verifies that the city of the branch and the city of the promotion are the same.

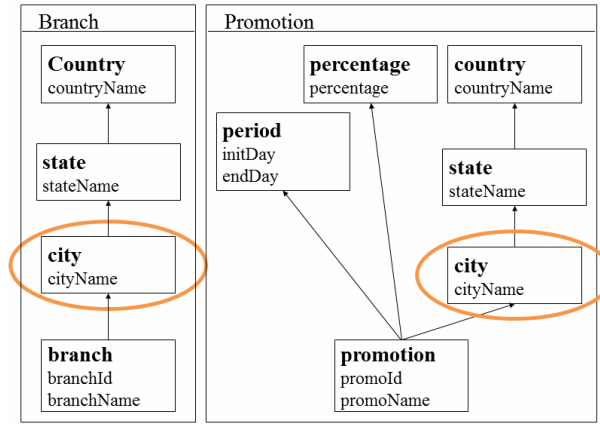


Figure 15: Dimensions and levels that give context to the fact table *Discounts*.

Table 5: DQ Metric *dwq_Example2*.

<i>Component</i>	DW
<i>DQ Dimension</i>	<i>Consistency</i>
<i>DQ Factor</i>	<i>Intra-relationship integrity</i> : It refers to how well integrity rules are met.
<i>DQ Metric</i>	Name: <i>dwq_Example2</i>
	Contextualizing object: DW dimensions
	Contextualized object: Fact table <i>Discounts</i>
	Granularity: Tuple
	Description: Check each tuple in the fact table <i>Discounts</i> . For each tuple in the fact table <i>Discounts</i> , it considers the promotion identifier <i>pi</i> and the branch identifier <i>bi</i> . Later, it takes of the dimension table <i>Promotion</i> the corresponding city <i>cp</i> for <i>pi</i> and of the dimension table <i>Branch</i> the corresponding city <i>cb</i> for <i>bi</i> . Finally, if <i>cp</i> = <i>cb</i> then it is a correct tuple. Otherwise, it is not a correct tuple.
	Result type: {0,1}

For example, we can see the following in Fig. 13:

- *tuple d1*: this tuple has the promotion *p1* and the branch identifier ‘28’. The promotion *p1* was created for the city ‘Florida’ as is shown in the dimension table *Promotion* in Fig 11, and the branch identifier ‘28’ belongs to the ‘Florida’ city as is shown in the dimension table *Branch* in Fig 10. The cities match, then the information in this tuple is consistent. Therefore, the tuple *d1* is correct.
- *tuple d2*: this tuple has the promotion *p2* and the branch identifier ‘31’. The promotion *p2* was created for the city ‘Colonia’ as is shown in the dimension table *Promotion* in Fig 11, and the branch identifier ‘31’ belongs to the ‘Rocha’ city as is shown in the dimension table *Branch* in Fig 10. The city of the promotion and the branch city in which the promotion was performed not match, then the information in this tuple is not consistent. Therefore, the tuple *d2* is incorrect.

As in the previous example, once all tuples in the fact table *Discounts* have been evaluated, the values obtained from the DQ metric *dmq_Example2* are saved in the metadata repository.

5.2 DQ in the DM

This section shows how analysis domain rules, which give context to each DM component, determine the quality in this component. We present two examples in which the domains considered are *Sales* and *Advertising*.

5.2.1 Example 3. Domain: Sales. Context: Domain Rule *RSales*

Table 6 shows, for the quality dimension *accuracy* and its quality factor *syntactic accuracy*, the quality assessment of the dimension table *Branch* with regard to the attribute *branchName*. This attribute represents the name of each branch. In this case, we consider the *Sales* domain in which there is a domain rule called *RSales* that says the following: “The name of each branch must contain the name of the supermarket, the name of the city to which it belongs the branch and the number of its identifier, in that order”. In this domain, this rule is necessary because the data for the Data Mart must be integrated with other data that verify this rule.

Table 6: DQ Metric *dmq_Example3*.

<i>Component</i>	DM
<i>Domain</i>	Sales
<i>Domain Rule</i>	<i>RSales</i> : The branch’s name structure must be <i>p1-p2-p3</i> , where <i>p1</i> is the supermarket’s name, <i>p2</i> is the city’s name to which it belongs and <i>p3</i> is the branch’s identifier.
<i>DQ Dimension</i>	<i>Accuracy</i>
<i>DQ Factor</i>	<i>Syntactic accuracy</i> : It refers to the syntactic accuracy of the data.
<i>DQ Metric</i>	Name: <i>dmq_Example3</i>
	Contextualizing object: Domain rule <i>RSales</i>
	Contextualized object: Dimension table <i>Branch</i>
	Granularity: Attribute
	Description: For each attribute <i>branchName bi</i> in each tuple of the dimension table <i>Branch</i> , the metric verifies that <i>bi</i> has the structure <i>p1-p2-p3</i> where <i>p1</i> must be <i>Super</i> (Supermarket’s name), <i>p2</i> must be the value in <i>cityName</i> (of the dimension table <i>Branch</i>) and <i>p3</i> the value in <i>branchId</i> (of the dimension table <i>Branch</i>). If the <i>branchName</i> attribute value verifies <i>RSales</i> the attribute is correct. Otherwise, it is incorrect.
	Result type: {0,1}

For example, in Fig. 10 corresponding to the dimension table *Branch*, we can see the following:

- *branch with id ‘29’*: the name of this branch is ‘Florida_29’, it is in the ‘Florida’ city (as shown in the column *cityName*) and the identifier of the branch is the number ‘29’, but the name of this branch does not contain the name of the supermarket: *Super*. Therefore, this tuple violates the rule *RSales*, so it is an incorrect tuple for the domain *Sales*.

- *branch with id '31'*: the name of this branch is 'Super_Rocha_31', the supermarket's name is *Super*, this branch belongs to the city 'Rocha' (as we can see in the column *cityName*) and the identifier of the branch is the number '31'. The name of this branch contains the information required by the rule *RSales*. Therefore, this tuple is correct for the domain *Sales*.

Once all tuples in the dimension table *Branch* have been evaluated, the values obtained from the DQ metric *dmq_Example3* are saved in the metadata repository. These metadata can later be used.

5.2.2 Example 4. Domain: Advertising. Context: Domain Rule RAdv

Table 7 shows, for the quality dimension *accuracy* and its quality factor *syntactic accuracy*, the quality assessment of the dimension table *Branch* with regard to the attribute *branchName*. This attribute represents the name of each branch. In this case, we consider the *Advertising* domain in which there is a domain rule called *RAdv* that says the following: "The name of each branch must contain the name of the city to which it belongs". In this domain, it is necessary to organize the promotions by cities. Therefore, this rule is necessary for organizing the marketing information.

Table 7: DQ Metric *dmq_Example4*.

<i>Component</i>	DM
<i>Domain</i>	Advertising
<i>Domain Rule</i>	<i>RAdv</i> : The name of each branch must contain the name of the city to which it belongs.
<i>DQ Dimension</i>	<i>Accuracy</i>
<i>DQ Factor</i>	<i>Syntactic accuracy</i> : It refers to the syntactic accuracy of the data.
<i>DQ Metric</i>	Name: <i>dmq_Example4</i>
	Contextualizing object: Domain rule <i>RAdv</i>
	Contextualized object: Dimension table <i>Branch</i>
	Granularity: Attribute
	Description: For each attribute <i>branchName</i> <i>bi</i> in each tuple of the dimension table <i>Branch</i> , the metric verifies that <i>bi</i> has the structure <i>p1-p2-p3</i> where <i>p2</i> must be the value in <i>cityName</i> (of the dimension table <i>Branch</i>), while <i>p1</i> and <i>p3</i> can have any value. If the <i>branchName</i> attribute value verifies <i>RAdv</i> the attribute is correct. Otherwise, it is incorrect.
	Result type: {0,1}

For example, in Fig. 10 corresponding to the dimension table *Branch*, we can see the following:

- *branch with id '25'*: the name of this branch is 'Supermarket'. The city name to which belongs this branch ('Maldonado', as is shown in the field *cityName*) is not in the name of the branch. Therefore, this tuple violates the domain rule *RAdv*, it is incorrect for the domain *Advertising*.
- *branch with id '29'*: the name of this branch is 'Florida_29'. The city name to which belongs this branch is 'Florida' (as is shown in the field *cityName*). The name of this branch contains the information required by the rule *RAdv*. Therefore, this tuple is correct for the domain *Advertising*.

It is important to highlight that tuples of the dimension table *Branch* that are incorrect for the domain *Sales* (eg branch with id '29'), are correct for the domain *Advertising*. This is so because different rules were applied to assess quality in two different domains. Therefore, we can see that different contexts (determined by domain rules), define different quality values. Once all tuples in the dimension table *Branch* have been evaluated, the values obtained from the DQ metric *dmq_Example4* are saved in the metadata repository. These metadata can later be used.

5.3 DQ in Use

In this section we show how the different elements that give context to the data presentation component determine the quality in this component. In the following we present two examples that take into account users from different domains: *Sales* and *Advertising*.

5.3.1 Example 5. Domain: Sales. Context: User profile

Table 8 presents the DQ metric *qiu_Example5*, for the quality dimension *freshness* and its quality factor *currency*, this last one refers to how updated are the data with regard to its source. The quality assessment is made on the fact table *Sales* (Fig. 12) with regard to the measure *salesAmount*. This DW measure represents the amount of sales made for each product in each branch in a day determined. In this case, we consider that the fact table *Sales* has a column *timestamp* that indicates the last update of the value of the measure *salesAmount*. The quality assessment is performed in two different contexts: with regard to a user sales general manager of the supermarket chain and with regard to a user sales manager of a branch. Therefore, for the quality assessment, it is necessary to consider the profile of each user according to his requirements:

- *user u₁*: since the user sales general manager of the supermarket chain needs to do different statistics, he requires that the last data update has been made, at most, the day before that the data are used. In other words, the data must be 24 hours old at most.
- *user u₂*: since the user sales manager of a branch needs to do different comparative analysis in real time, he requires that the last data update has been made, at most, one hour before that the data are used. In other words, the data must be 1 hour old at most.

Table 8: DQ Metric *qiu_Example5*.

<i>Component</i>	Data presentation layer
<i>User u₁</i>	<i>Role</i> : Sales general manager. <i>Quality requirement</i> : Data must be 24 hours old at most.
<i>User u₂</i>	<i>Role</i> : Sales manager of a branch. <i>Quality requirement</i> : Data must be 1 hours old at most.
<i>DQ Dimension</i>	<i>Freshness</i>
<i>DQ Factor</i>	<i>Currency</i> : It refers to how updated are the data with regard to its source.
<i>DQ Metric</i>	<i>Name</i> : <i>qiu_Example5</i> <i>Contextualizing object</i> : User profile. <i>Contextualized object</i> : Fact table <i>Sales</i> <i>Granularity</i> : Table <i>Description</i> : It gets the latest update of the fact table <i>Sales</i> <i>tu</i> is the corresponding value of the field <i>timestamp</i> in the latest update and <i>tp</i> is the corresponding value at the present moment. If $tp - tu \leq 24$ hs. then it is a correct tuple for the user <i>u₁</i> . Otherwise, it is not a correct tuple for the user <i>u₁</i> . If $tp - tu \leq 1$ hs. then it is a correct tuple for the user <i>u₂</i> . Otherwise, it is not a correct tuple for the user <i>u₂</i> . <i>Result type</i> : {0,1}

It is possible to see that the quality requirement of the user *u₂* is more demanding than the quality requirement of the user *u₁*. Therefore, data that are valid for the user *u₂* are valid for the user *u₁*. However, data that are valid for the user *u₁* are not valid for the user *u₂*. In this case, although is a quality requirement which has influence over the outcome of the DQ metric *qiu_Example5*, is the user profile who determines which of the two requirements will be taken into account. Thus, the user profile is who determines the outcome obtained from the DQ metric *qiu_Example5*, so the user is who determines the context in which the quality assessment is performed.

5.3.2 Example 6. Domain: Advertising. Context: Quality requirement

Table 9 presents the DQ metric *qiu_Example6*, for the quality dimension *accuracy* and its quality factor *syntactic accuracy*, this last one refers to the syntactic accuracy of the data. The quality assessment is made on the dimension table *Branch* (Fig. 10). In this case, the context is determined by a quality requirement. Such a requirement is requested by the advertising manager. This user belongs to the *Advertising* domain. The quality requirement of the user says that the 100% of the branches must have the name written correctly.

For the quality assessment, the DQ metric *qiu_Example6* requests the quality values obtained in the execution of the DQ metric *qiu_Example4*. Last one, applied in the DM component (see Table 7).

Table 9: DQ Metric *qiu_Example6*.

<i>Component</i>	Data presentation layer
<i>User</i>	Role: Advertising manager
	Quality requirement: 100% of the branches must have the name written correctly.
<i>DQ Dimension</i>	Accuracy
<i>DQ Factor</i>	Syntactic accuracy: It refers to the syntactic accuracy of the data.
<i>DQ Metric</i>	Name: <i>qiu_Example6</i>
	Contextualizing object: Quality requirement
	Contextualized object: Dimension table <i>Branch</i>
	Granularity: Table
	Description: This DQ metric asks the results obtained in DQ metric <i>dmq_Example4</i> . The dimension table <i>Branch</i> is correct if all tuples returned by DQ metric <i>dmq_Example4</i> are correct. Otherwise, the dimension table <i>Branch</i> is incorrect.
	Result type: {0,1}

5.4 Case Study Summary

In this section, we present in the Table 10 a summary with all the examples developed before. In the first column the DWS component in which the quality assessment was performed appears. Following, the number of the example is presented. Later the contextualized and contextualizing objects are shown, and finally the quality dimension and the quality factor assessed appear.

Table 10: Summary Case Study

Component	Example	Contextualized object	Contextualizing object	Dimension/Factor
DW	1	Dimension table <i>Branch</i>	Organization's document	Accuracy/Semantic accuracy
	2	Fact table <i>Discounts</i>	DW dimensions	Consistency/ Intra-relationship integrity
DM	3	Dimension table <i>Branch</i>	Domain rule (<i>RSales</i>) over the name of the branches	Accuracy/Syntactic accuracy
	4	Dimension table <i>Branch</i>	Domain rule (<i>RAdv</i>) over the name of the branches	Accuracy/Syntactic accuracy
Data presentation layer	5	Fact table <i>Sales</i>	User profile	Freshness/Currency accuracy
	6	Dimension table <i>Branch</i>	Quality requirement	Accuracy/Syntactic accuracy

6 Conclusions and Future Work

This work proposes to assess DQ in DWS taking the context into account. This context varies depending on the DWS component considered. Therefore DQ assessment must be defined for each component, i.e. DQ metrics must be defined based on the context of each component. In this work, contexts and DQ are defined for three components of DWS: DW, DMs and data presentation layer.

For the definition of DQ at the different DWS components, this work is supported by two quality approaches: Meeting Requirements and Fitness for Use. The former is applied for data quality assessment in two components, DW and DM, emphasizing on compliance with the requirements. The latter is applied for data quality assessment in the data presentation layer, emphasizing on meeting the user needs.

In addition, this work presents a literature review that is built over the research question *How contexts may be used for assessing DQ in DW?*, which is split into other three research questions. The result of the review is commented organized by the posed questions, and besides, some specific results regarding the different approaches

of contexts are presented. As shown in this paper, there is no consensus in the literature regarding the management of the context for the data quality assessment in DWS.

To show how contexts are used in quality assessment, we propose a case study in which DWS components are considered and DQ metrics are defined for them. Once the component to be studied is selected, the quality dimension and quality factor to be assessed are chosen. Then, the context and the context-based quality metric are defined. Different contexts may determine different quality values for the same metric.

While DQ assessment in DWS is not a new problem, since there is a lot of literature that investigates and raises the need for managing DQ in this type of systems, no proposals are found with the approach presented by this research. This first proposal defines a framework where DQ problems are placed in DWS and serves as a basis for the integral quality management (assessment, improvement, maintenance, etc.). Therefore, this research is the starting point for new research addressing new challenges.

The immediate challenge, around the proposal presented above, is to define a formal model of the system, of the contexts and quality defined for the DWS components. This will allow detailed and deep specification of the proposed solution. On the other hand, it is interesting to analyse the relationship between DQ and quality in use, addressed by the Meeting Requirements and Fitness for Use approaches, respectively.

References

- [1] M. Scannapieco and T. Catarci. Data quality under a computer science perspective. *Archivi & Computer*, 2:1-15, 2002.
- [2] M. Scannapieco, P. Missier and C. Batini. Data quality at a glance. *Datenbank-Spektrum*, 14:6-14, 2005.
- [3] D.M. Strong, Y.W. Lee and R.Y. Wang. Data quality in context. *Commun. ACM*, 40(5):103-110, May 1997. [Online]. Available: <https://doi.org/10.1145/253769.253804>
- [4] J. Juran and A. Blanton Godfrey. *Quality handbook*. Republished McGraw-Hill, 1999.
- [5] C. Bolchini, C. A. Curino, G. Orsi, E. Quintarelli, R. Rossato, F. A. Schreiber and L. Tanca. And what can context do for data? *Commun. ACM*, 52(11):136-140, November 2009. [Online]. Available: <https://doi.org/10.1145/1592761.1592793>
- [6] K. Henriksen, J. Indulska and A. Rakotonirainy. Modeling context information in pervasive computing systems. In *Proceedings of the First International Conference on Pervasive Computing*, Pervasive '02, pages 167-180, London, UK, UK, 2002. Springer-Verlag. [Online]. Available: https://doi.org/10.1007/3-540-45866-2_14
- [7] F. Serra and A. Marotta. Data Quality in Data Warehouse Systems: A context-based approach. In *IEEE XLII Latin American Computing Conference (CLEI 2016)*. Valparaíso, Chile. [Online]. Available: <https://doi.org/10.1109/clei.2016.7833371>
- [8] J. E. Olson. *Data quality: the accuracy dimension*. Morgan Kaufmann, 2003.
- [9] R. Y. Wang and D.M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5-33, March 1996. [Online]. Available: <https://doi.org/10.1080/07421222.1996.11518099>
- [10] M. Pamela Neely. The product approach to data quality and Fitness for use: a framework for analysis. *Proc. of 10th International Conference on Information Quality*, 2005.
- [11] A. D. Chapman. *Principles of data quality*. GBIF, 2005.
- [12] M. Golfarelli and S. Rizzi. *Data Warehouse design: Modern principles and methodologies*. McGraw-Hill, Inc., 2009.
- [13] W. H. Inmon. *Building the data warehouse*. John Wiley & sons, 2005.
- [14] E. Malinowski and E. Zimányi. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications (Data-Centric Systems and Applications)*. Springer Publishing Company, Incorporated, 1 edition, 2008.
- [15] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley, 2nd edition, April 2002.

- [16] P. Ciaccia and R. Torlone. Modeling the propagation of user preferences. In Manfred Jeusfeld, Lois Delcambre, and Tok-Wang Ling, editors, *Conceptual Modeling ER 2011*, volume 6998 of *Lecture Notes in Computer Science*, pages 304-317. Springer Berlin Heidelberg, 2011.
- [17] M. Bazire and P. Brézillon. Understanding context before using it. In *Modeling and using context*, pages 29-40. Springer, 2005. [Online]. Available: https://doi.org/10.1007/11508373_3
- [18] N. Prat and S. Madnick. Measuring Data Believability: A Provenance Approach. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 393-393, January 2008. [Online]. Available: <https://doi.org/10.1109/hicss.2008.243>
- [19] A. Rai, V. Dubey, K. K. Chaturvedi and P. K. Malhotra. Design and development of data mart for animal resources. *Computers and Electronics in Agriculture*, 64(2):111-119, 2008. [Online]. Available: <https://doi.org/10.1016/j.compag.2008.04.009>
- [20] J. Rodic and M. Baranovic. Generating Data Quality Rules and Integration into ETL Process. In *Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP, DOLAP '09*, pages 65-72, New York, NY, USA, 2009. ACM. [Online]. Available: <https://doi.org/10.1145/1651291.1651303>
- [21] K. Ali and M.A. Warraich. A framework to implement data cleaning in enterprise data warehouse for robust data quality. In *Information and Emerging Technologies (ICIET), 2010 International Conference on*, pages 1-6, 2010. [Online]. Available: <https://doi.org/10.1109/iciet.2010.5625701>
- [22] V. Santos and O. Belo. Modeling ETL Data Quality Enforcement Tasks Using Relational Algebra Operators. *Procedia Technology*, 9(0):442-450, 2013. *CENTERIS 2013 - Conference on Enterprise Information Systems. International Conference on Project Management/HCIIST 2013 - International Conference on Health and Social Care Information Systems and Technologies*. [Online]. Available: <https://doi.org/10.1016/j.protcy.2013.12.049>
- [23] F. Daniel, F. Casati, T. Palpanas, and O. Chayka. Managing data quality in business intelligence applications. In *QDB/MUD*, pages 133-143, 2008.
- [24] B. Wieder and M.L. Ossimitz. The impact of business intelligence on the quality of decision making a mediation model. *Procedia Computer Science*, 64:1163-1171, 2015. *Conference on Enterprise Information Systems/International Conference on Project Management/Conference on Health and Social Care Information Systems and Technologies*, October 7-9, 2015. [Online]. Available: <https://doi.org/10.1016/j.procs.2015.08.599>
- [25] W. Gongora de Almeida, R.T. de Sousa, F.E. de Deus, G.D. Amvame Nze, and F.L. Lopes de Mendonca. Taxonomy of data quality problems in multidimensional Data Warehouse models. In *Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on*, pages 1-7, June 2013.
- [26] L. Oukid, O. Asfari, F. Bentayeb, N. Benblidia and O. Boussaid. CXT-cube: Contextual Text Cube Model and Aggregation Operator for Text OLAP. In *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, DOLAP '13*, pages 27-32, New York, NY, USA, 2013. ACM. [Online]. Available: <https://doi.org/10.1145/2513190.2513201>
- [27] B. Vela, J. N. Mazn, C. Blanco, E. Fernndez-Medina, J. Trujillo, and E. Marcos. Development of secure xml data warehouses with qvt. *Information and Software Technology*, 55(9):1651-1677, 2013. [Online]. Available: <https://doi.org/10.1016/j.infsof.2013.03.003>
- [28] R. Thollot, F. Brauer, W. M. Barczynski, and M.A. Aufaure. Text-to-query: Dynamically Building Structured Analytics to Illustrate Textual Content. In *Proceedings of the 2010 EDBT/ICDT Workshops, EDBT '10*, pages 14:1-14:8, New York, NY, USA, 2010. ACM.
- [29] J.M. Perez, R. Berlanga, M.J. Aramburu, and T.B. Pedersen. Towards a Data Warehouse Contextualized with Web Opinions. In *e-Business Engineering, 2008. ICEBE '08. IEEE International Conference on*, pages 697-702, October 2008. [Online]. Available: <https://doi.org/10.1109/icebe.2008.43>
- [30] H. Tahir and P. Brezillon. A shared context approach for supporting experts in data ETL (Extraction, Transformation and Loading) processes. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 720-725, November 2011.
- [31] N. Prat, I. Comyn-Wattiau, and J. Akoka. Combining objects with rules to represent aggregation knowledge in data warehouse and OLAP systems. *Data & Knowledge Engineering*, 70(8):732-752, 2011. *Pushing Artificial*

Intelligence in Database and Data Warehouse Systems. [Online]. Available: <https://doi.org/10.1016/j.datak.2011.03.004>

- [32] S. Aissi, T. Sboui, M.S. Gouider, M.A. Ben Hassine and L. Ben Said. A recommendation approach to enhance the interoperability between spatial datacubes. *Procedia Computer Science*, 56:558-565, 2015. The 10th International Conference on Future Networks and Communications (FNC 2015). The 12th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2015) Affiliated Workshops. [Online]. Available: <https://doi.org/10.1016/j.procs.2015.07.254>
- [33] M. Mehdi, R. Sahay, W. Derguech and E. Curry. On-the-Fly Generation of Multidimensional Data Cubes for Web of Things. In *Proceedings of the 17th International Database Engineering & Applications Symposium, IDEAS'13*, pages 28-37, New York, NY, USA, 2013. ACM. [Online]. Available: <https://doi.org/10.1145/2513591.2513655>
- [34] V.E. Silva Souza, J.N. Mazón, I. Garrigós, J. Trujillo and J. Mylopoulos. Monitoring Strategic Goals in Data Warehouses with Awareness Requirements. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC'12*, pages 1075-1082, New York, NY, USA, 2012. ACM. [Online]. Available: <https://doi.org/10.1145/2245276.2231944>
- [35] X. Franch, A. Mate, J.C. Trujillo and C. Cares. On the joint use of i* with other modelling frameworks: A vision paper. In *Requirements Engineering Conference (RE)*, 2011 19th IEEE International, pages 133-142, August 2011.
- [36] P. Hernandez, I. Garrigos and J. Mazón. Modeling Web Logs to Enhance the Analysis of Web Usage Data. In *Database and Expert Systems Applications (DEXA)*, 2010 Workshop on, pages 297-301, August 2010.
- [37] O. Badreddin, G. Mussbacher, D. Amyot, S.A. Behnam, R. Rashidi-Tabrizi, E. Braun, M. Alhaj and G. Richards. Regulation-Based Dimensional Modeling for Regulatory Intelligence. In *Requirements Engineering and Law (RELAW)*, 2013 Sixth International Workshop on, pages 1-10, July 2013. [Online]. Available: <https://doi.org/10.1109/relaw.2013.6671340>
- [38] O. Glorio, J.N. Mazón, I. Garrigós and J. Trujillo. A personalization process for spatial data warehouse development. *Decision Support Systems*, 52(4):884-898, 2012. 1)Decision Support Systems for Logistics and Supply Chain Management 2)Business Intelligence and the Web.
- [39] A. Chen, V.M. Markowitz, E. Szeto, K. Palaniappan and K. Chu. Maintaining a microbial genome & metagenome data analysis system in an academic setting. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management, SSDBM'14*, pages 3:1-3:11, New York, NY, USA, 2014. ACM. [Online]. Available: <https://doi.org/10.1145/2618243.2618244>
- [40] P. Stack. Development of a Mobile Platform to Support Building Maintenance Engineering. In *Computer Software and Applications Conference Workshops (COMPSACW)*, 2012 IEEE 36th Annual, pages 482-487, July 2012. [Online]. Available: <https://doi.org/10.1109/compsacw.2012.91>
- [41] L. Oukid, O. Asfari, F. Bentayeb, N. Benblidia and O. Boussaid. CXT-cube: Contextual Text Cube Model and Aggregation Operator for Text OLAP. In *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, DOLAP '13*, pages 27-32, New York, NY, USA, 2013. ACM. [Online]. Available: <https://doi.org/10.1145/2513190.2513201>
- [42] R. Nachida and N. Fahima. A contextual personalized recommender system for mobile OLAP. In *Information Technology and e-Services (ICITeS)*, 2012 International Conference on, pages 1-7, March 2012. [Online]. Available: <https://doi.org/10.1109/icites.2012.6216674>
- [43] R. Thollot, F. Brauer, W.M. Barczynski and M.A. Aufaure. Text-to-query: Dynamically Building Structured Analytics to Illustrate Textual Content. In *Proceedings of the 2010 EDBT/ICDT Workshops, EDBT'10*, pages 14:1-14:8, New York, NY, USA, 2010. ACM.
- [44] M.M. Hamad and A.A. Jihad. An Enhanced Technique to Clean Data in the Data Warehouse. In *Developments in E-systems Engineering (DeSE)*, 2011, pages 306-311, December 2011. [Online]. Available: <https://doi.org/10.1109/dese.2011.32>
- [45] M. Munawar, N. Salim, and R. Ibrahim. Towards Data Quality into the DataWarehouse Development. In *Dependable, Autonomic and Secure Computing (DASC)*, 2011 IEEE Ninth International Conference on, pages 1199-1206, December 2011.

- [46] M. Helfert and O. Foley. A Context Aware Information Quality Framework. In Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09. Fourth International Conference on, pages 187-193, November 2009. [Online]. Available: <https://doi.org/10.1109/coinfo.2009.65>
- [47] H.T. Moges, K. Dejaeger, W. Lemahieu and B. Baesens. A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management*, 50(1):43-58, 2013. [Online]. Available: <https://doi.org/10.1016/j.im.2012.10.001>
- [48] D.S. Alberts, M. Vassiliou, and J. Agre. C2 information quality: An enterprise systems perspective. In MILITARY COMMUNICATIONS CONFERENCE, 2012 - MILCOM 2012, pages 1-7, October 2012. [Online]. Available: <https://doi.org/10.1109/milcom.2012.6415625>
- [49] G. Rogova, M. Hadzagic, M. St-Hilaire, M.C. Florea, and P. Valin. Contextbased information quality for sequential decision making. In Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2013 IEEE International Multi-Disciplinary Conference on, pages 16-21, February 2013.
- [50] B. Catania, G. Guerrini, A. Belussi, F. Mandreoli, R. Martoglia and W. Penzo. Wearable Queries: Adapting Common Retrieval Needs to Data and Users. In Proceedings of the 7th International Workshop on Ranking in Databases, DBRank '13, pages 7:1-7:3, New York, NY, USA, 2013. ACM. [Online]. Available: <https://doi.org/10.1145/2524828.2524835>
- [51] A.L. McNab and D.A. Ladd. Information Quality: The Importance of Context and Trade-Offs. In System Sciences (HICSS), 2014 47th Hawaii International Conference on, pages 3525-3532, January 2014.
- [52] G. Rogova, M. Hadzagic, M. St-Hilaire, M.C. Florea and P. Valin. Contextbased information quality for sequential decision making. In Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2013 IEEE International Multi-Disciplinary Conference on, pages 16-21, February 2013.
- [53] R. Price and G. Shanks. DQ Tags and Decision-Making. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pages 1-10, January 2010. [Online]. Available: <https://doi.org/10.1109/hicss.2010.159>
- [54] A. Manzoor, H.L. Truong and S. Dustdar. Quality Aware Context Information Aggregation System for Pervasive Environments. In Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on, pages 266-271, May 2009. [Online]. Available: <https://doi.org/10.1109/waina.2009.144>
- [55] D. Zheng, J. Wang and B. Kerong. Evaluation of Quality Measure Factors for the Middleware Based Context-Aware Applications. In Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on, pages 403-408, May 2012. [Online]. Available: <https://doi.org/10.1109/icis.2012.46>
- [56] T. Dasu, J.M. Loh and D. Srivastava. Empirical glitch explanations. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pages 572-581, New York, NY, USA, 2014. ACM. [Online]. Available: <https://doi.org/10.1145/2623330.2623716>
- [57] J. Han, D. Jiang and L. Li. Automatic accuracy assessment via hashing in multiple-source environment. *Expert Systems with Applications*, 37(3):2609-2620, 2010. [Online]. Available: <https://doi.org/10.1016/j.eswa.2009.08.023>
- [58] A. Sundararaman. A framework for linking Data Quality to business objectives in decision support systems. In Trendz in Information Sciences and Computing (TISC), 2011 3rd International Conference on, pages 177-181, December 2011. [Online]. Available: <https://doi.org/10.1109/tisc.2011.6169110>
- [59] M. Milani, L. Bertossi and S. Ariyan. Extending contexts with ontologies for multidimensional data quality assessment. In Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on, pages 242-247, March 2014. [Online]. Available: <https://doi.org/10.1109/icdew.2014.6818333>
- [60] R.G. Tiwari, M. Husain, B. Gupta and A. Agrawal. Amalgamating Contextual Information into Recommender System. In Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on, pages 15-20, November 2010. [Online]. Available: <https://doi.org/10.1109/icetet.2010.110>
- [61] M. Goller and S. Berger. Slowly Changing Measures. In Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, DOLAP '13, pages 47-54, New York, NY, USA, 2013. ACM. [Online]. Available: <https://doi.org/10.1145/2513190.2513194>

- [62] O. Isk, M.C. Jones and A. Sidorova. Business intelligence success: The roles of BI capabilities and decision environments. *Information & Management*, 50(1):13-23, 2013. [Online]. Available: <https://doi.org/10.1016/j.im.2012.12.001>
- [63] Z. Huang and H. Peng. Improving Uncertain Data-Quality through Effective Use of Knowledge Base. In *Wireless Communications, Networking and Mobile Computing*, 2008. WiCOM '08. 4th International Conference on, pages 1-4, October 2008. [Online]. Available: <https://doi.org/10.1109/wicom.2008.2517>
- [64] J. Han, D. Jiang and L. Li. Automatic accuracy assessment via hashing in multiple-source environment. *Expert Systems with Applications*, 37(3):2609-2620, 2010. [Online]. Available: <https://doi.org/10.1016/j.eswa.2009.08.023>
- [65] P. B. Crosby. *Quality Is Free. The art of making quality certain*. New York: New American Library, 17, 1979.
- [66] R. W. Hoyer and B. BY Hoyer. What is quality. *Quality Progress*, 34(7):53-62, 2001.
- [67] ISO/IEC 25010:2011. *Systems, software Quality Requirements, Evaluation*. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35733. Last access in November, 2016
- [68] M. Herrera, M.A. Moraga, I. Caballero, and C. Calero. Quality in use model for web portals (qiuwep). In Florian Daniel and FedericoMichele Facca, editors, *Current Trends in Web Engineering*, volume 6385 of *Lecture Notes in Computer Science*, pages 91-101. Springer Berlin Heidelberg, 2010.
- [69] F. Carpani and R. Ruggia. An Integrity Constraints Language for a Conceptual Multidimensional Data Model. In *Proceedings of the Thirteenth International Conference on Software Engineering & Knowledge Engineering (SEKE'2001)*. Buenos Aires, Argentina, June 13-15, 2001, 2001, pp. 220–227.