Context based Data Quality Rules for Multidimensional Data

Camila Sanz[1]

[1]*Instituto de Computación, Facultad de Ingeniería, Universidad de la República*
Supervised by Adriana Marotta

Abstract

Data quality evaluation and improvement is an important asset in every system, particularly in systems which aim is to analyse data, such as those that are based on multidimensional data models. When talking about data quality the main approach found in literature is *fitness for use* which means that data quality cannot be evaluated nor improved without taking context into account. Evaluating data quality over systems with a multidimensional model is clearly context dependent. However, there is not enough generality in the solutions found in the literature for context-based data quality management, which means that for every particular case the problem needs to be redefined. In this PhD proposal we aim to reach to a formal definition of every concept mentioned above and their interactions. As a result it would be possible that, given a particular multidimensional model and its context, a set of Data Quality rules can be generated in a simple way.

Keywords

data quality, multidimensional data, context

1. Introduction

Data Quality (DQ) is a multifaceted concept, since there are many aspects that can be taken into account when trying to define and measure the quality of data. These aspects are called DQ dimensions, while DQ metrics are defined in order to measure them [1].

Data Warehouse (DW) systems are decision-oriented information systems and as so, a fundamental asset for decision making. DWs are populated with data extracted from heterogeneous sources which is transformed to be queried and analyzed with a multidimensional perspective, allowing aggregations by different criteria. Multidimensional data model is typically used for designing DWs and for doing analysis on top of them. The main concepts of this model are dimensions, hierarchies, facts and cubes, also including as an essential tool, a set of multidimensional operations that allow navigating and aggregating data.

In these systems DQ is an unavoidable issue, since it is compromised at different moments of the DW lifecycle, such as ETL and multidimensional operations. Specific DQ problems appear due to multidimensional model characteristics (described above). DQ management allows DQ improvement when it is possible, and also DQ awareness by the user, ensuring decision making is not biased by poor quality data.

There is consensus in the literature about the importance of considering context in DQ management. The well-known *fitness for use* approach has been widely adopted [1], accepting that DQ cannot be evaluated nor improved ignoring the information about the context where data will be used. In the case of DWs, context can be useful for compensating missing data, correcting errors, detecting inconsistencies, and many other quality-related tasks.

As DQ is context dependent, DQ dimensions and metrics are specific for each domain and use case, therefore, solutions are highly dependent on each particular case. Formalization is needed to provide an abstraction level that gives generality to solutions, allowing the instantiation of them for each particular case.

Although there has been certain progress in research about context-oriented DQ for DW, we believe that there is still a deep gap for arriving to well-formalized integral and robust solutions. There are few works that propose formalizations for these concepts, and they do not address DQ as an integral discipline, including DQ dimensions and metrics management, and differentiating the tasks in DQ management, mainly evaluation and improvement.

This work is a step forward the formalization of DW, DQ and context, in a general way, so that it allows managing context-oriented DQ in DW for particular cases, in a robust and systematic way. Among DQ dimensions, we focus on *consistency*, *accuracy* and *completeness* [1], as they illustrate very common DW quality problems.

The rest of the document is structured as follows: in section 2 we mention some works related to DW, DQ and context focusing on existing formalizations, in section 3 we present the PhD proposal in terms of the problem and solution approach, and in section 4 we conclude and mention the next steps to be followed.

csanz@fing.edu.uy (C. Sanz)

2. Related Work

As said before, DQ is context dependent [2, 3, 4, 5, 6, 7], as it is perceived differently according to the application domain of the data, the user or even the location in which it is being used. For this reason, the context becomes an essential part of DQ definition. On its own, data context is an ambiguous concept and in general it is specifically defined for each particular application. Commonly, it refers to user and location aspects [8, 9, 4, 10], but many other aspects can be considered for its definition.

What is left of this section is focused on the formalization of context and of the solutions for DQ over Multidimensional Data using context, as these are the main aspects addressed in our work.

Context Formalization. Considering works that propose formal models for contexts, two interesting approaches were found: using ontologies [11, 12, 13, 14, 15] and using first order predicates [16].

When using ontologies, some works present formal specifications that are absolutely domain dependant. In [12], the proposal of an access control mechanism is constructed by the context and the user profile, both modeled using ontologies. The main drawback of this approach is that it only considers the user context and that it is proposed for a specific domain. In [11] the context is specified in a more general way. The authors identify components that may belong to any context: people, activity, location and computational entity. Each specific domain is modeled with a particular ontology that is merged with the identified components. A similar approach is presented in [15], where a context model is presented considering different elements that should be present, such as the local context or the surrounding context. Each of these concepts are later mapped to domain ontologies in order to contextualize DWs. With the idea of mappings, in [14] domain ontologies are formalized and used in order to give context for a particular user that is modeled using an ontology. To do so, a formal mapping between both ontologies is proposed. Finally, with the aim of obtaining context from an ontology, in [13] a mathematical model is proposed. The authors' approach is to calculate the distance between different concepts in an ontology in order to determine the context of particular data.

In [16] first order predicates are used to formalize the context. However, the authors do not present a general context formalization that can be instantiated for particular cases. As in [12], the main problem with the approach is the lack of generality.

Context-based DQ over Multidimensional Data Existing work about contexts, DW and DQ in general, is analyzed in [17], where an exhaustive literature review is presented. The authors show that, although there are many works that consider the context for DQ, and many that consider the context for DW, very few address the problem of managing DQ in DW considering the context (i.e., relating the three issues simultaneously).

In [18] a model in which DQ is addressed at the ETL process is presented. Domain ontologies are used to model the process and business rules are mapped to those ontologies through different quality metrics. Even though it is not explicitly said, both ontologies and business rules are considered as context.

Both [19, 20] are based on [21], where Hurtado-Mendelzon's multidimensional model for DW is presented, making minor adaptations to it, according to the specific needs of each work. Even if the main goal is, in both of them, context based DQ evaluation, the use of the multidimensional data model differs.

In [19], the multidimensional model is used to model a part of the context that includes relationships between its components. To give context to relational databases, the authors adapt the model of [21] to a relational schema and combine it with quality predicates.

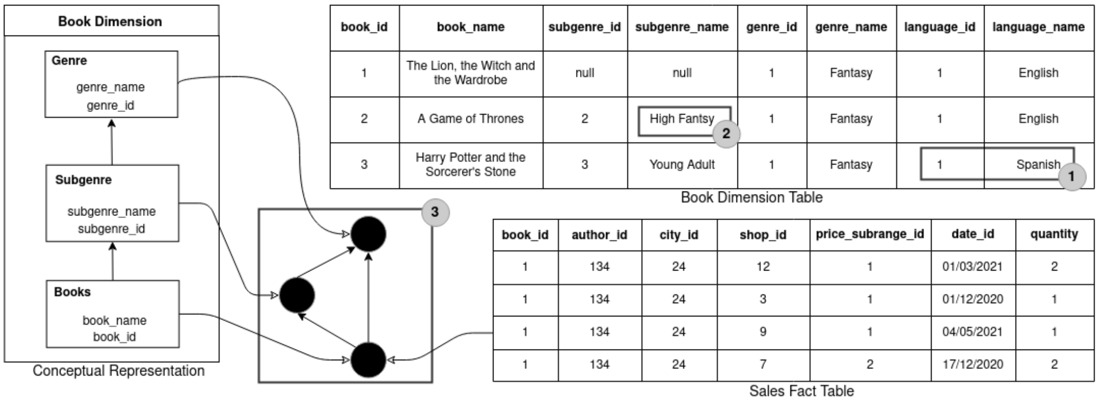
On the other hand, in [20] quality evaluation over a multidimensional model is specified with logical rules that includes context. In this case, the specification presented in [21] is adapted to be used in the rules definition.

Although there is a lot of work done in order to formalize the context for a dataset, there are two main aspects that remain unsolved: the context formalizations are not general enough so that they can be instantiated for the different particular cases, and there is very little work on formally specifying the context for DQ management in DW. Our work proposes a general formalization of a DW and its context that enables the instantiation for any particular case, and on top of this, it proposes the definition and formalization of context aware DQ rules for evaluation and improvement of the DW quality.

Our work share some aspects with many of the presented above. We particularly inspire on the multidimensional model proposed in [19] and the quality rules idea presented in [20]. We find specially relevant the mappings ideas presented in [14, 15] and the idea of determining the context of a dataset within a particular ontology presented in [13].

3. Thesis Proposal

This section presents the research problem addressed by this work and the solution approach. First, we illustrate the problem with an example, then we state the problems to solve in a general way and finally, we present the main aspects of our approach, through specific parts of the solution and examples of our proposal, trying to cover the whole picture of the proposed solution.



book_id	book_name	subgenre_id	subgenre_name	genre_id	genre_name	language_id	language_name
1	The Lion, the Witch and the Wardrobe	null	null	1	Fantasy	1	English
2	A Game of Thrones	2	High Fantsy	1	Fantasy	1	English
3	Harry Potter and the Sorcerer's Stone	3	Young Adult	1	Fantasy	1	Spanish

book_id	author_id	city_id	shop_id	price_subrange_id	date_id	quantity
1	134	24	12	1	01/03/2021	2
1	134	24	3	1	01/12/2020	1
1	134	24	9	1	04/05/2021	1
1	134	24	7	2	17/12/2020	2

Figure 1: Data Quality Problems

3.1. Running Example

DQ in DW systems involves typical DQ management over data attributes, but also includes problems related to multidimensional operations results. To illustrate both type of problems we use an example, whose main concepts are shown in Figure 1.

The example refers to a Sales DW, implemented in a relational star schema, which consists of a fact table *Sales*, related to many dimension tables with data about books, authors, cities, dates, etc. Figure 1 shows *Sales Fact Table* and *Book Dimension Table*. Additionally, the figure presents the conceptual representation of one hierarchy of Book dimension. This hierarchy is composed by three categories, named *Books*, *Subgenre* and *Genre*. We consider all hierarchies to be homogeneous [21], i.e., every member from a category has exactly one parent in the category above.

Different DQ problems may arise in this system, such as the ones showed in Figure 1:

DQ problems in attributes data. Rectangle 1 shows both an inconsistency between the attributes `language_id` and `language_name` and also a semantic accuracy problem because “Harry Potter and the Sorcerer’s Stone” is written in English. In rectangle 2 a syntactic accuracy problem is presented, it should say “High Fantasy” instead of “High Fantsy” .

Summarizability problem. Rectangle 3 shows summarizability problem [22] over *Book dimension*. Ideally, a roll-up from *Book* to *Genre* and the composition of the roll-up operations from *Book* to *Subgenre* and from *Subgenre* to *Genre* should return the same result. However, due to a DQ problem they may not return the same result. When looking at *Book Dimension Table*, book with id 1 does not have a value in the `subgenre` attribute. This means that a roll-up from *Book* to *Subgenre* will loose information and some sales in *Sales Fact Table* will not be considered when the roll-up from *Subgenre* to *Genre* takes place. However, when the roll-up is done directly from *Book* to *Genre*, no information is lost.

3.2. Research Problem

The research problem addressed by this work is the definition of formal rules for DQ assessment and improvement for a DW, taking the context into account.

In order to tackle this problem, we state the following sub-problems to be solved: (i) formal definitions for both DW and context, which allow the instantiation of any particular DW or context, (ii) definition of a mechanism for the interaction between DW and context, enabling the use of different formal languages to represent each one, (iii) definition and formalization of DQ assessment and improvement rules for the DQ dimensions: *accuracy*, *consistency* and *completeness* [1], and (iv) solution implementation, which integrates all the components in a unique system.

In order to test and validate the solution, a real use case consisting of a particular DW and its context should be designed and implemented. Afterwards, metrics and cleaning tasks for consistency, accuracy and completeness, should be implemented, as instantiations of the proposed DQ rules. Finally, we should carry out a comparison between the obtained results with our solution and results obtained with an analogous solution that does not consider context in DQ evaluation and improvement.

3.3. Approach

We use the formalization presented by Hurtado and Mendelzon [21] to formalize the DW, making some minor extensions and modifications in order to adapt the model to our goals.

The context is modeled through domain ontologies: given an OWL ontology O, we consider its classes named $Cl = \{Cl_1, \dots Cl_c\}$; its object properties named $OP = \{OP_1 \dots OP_{op}\}$, where $dom(OP_j)$ and $range(OP_j)$ are its domain and range; and its data properties $DP = \{DP_1, \dots DP_{dp}\}$, where $dom(DP_j)$ is a class and $range(DP_j)$ is a data type.

Mappings are defined as a mechanism for the interaction between DW and context (issue (iii) of previous section). They are ternary relations, where the first argument is the DW element, the second argument is the ontology element and the third is a Boolean that indicates if the mapping is total, which means that both the element of the DW and the context represent the same real world entity. We introduce as an example the definition of mappings for Dimensions and Categories.

Dimensions: $MapDim \subseteq \{\mathcal{S}[1], \dots, S[n]\} \times Cl \times \{true, false\}$ maps DW dimensions, using the specification taken from [21], to ontology classes.

Categories: $MapCat \subseteq (\mathscr{C}_1 \cup \dots \cup \mathscr{C}_n) \times (Cl \cup DP) \times \{true, false\}$ maps DW categories, using the specification from [21], either to ontology classes or to ontology data properties. If a category is mapped to a data property dp, then there must exist a mapping between either the dimension to which the category belongs or another related category of the same dimension, and the class $dom(dp)$.

Running Example The ontologies chosen to give context to Book Dimension presented in section 3.1 are "The British National Library"[1] ontology and "The Book Vocabulary Metadata"[2] both ontologies represent information about books and other aspects related to them such as authors or languages.

Figure 2 shows the ontologies that are mapped to Book dimension. For example from "British National Library" ontology we map *Book* category to `bibo:Book` class, this mapping is formalized as $MapCat(book, bibo : Book, true)$. In this case the mapping is total because both the category *Book* and the ontology class `bibo:Book` represent a book in the real world. This connection between both ontologies is represented in Figure 2 by the dotted line.

Mappings are fundamentally used to define the context of interest. Once the DW elements are located in the chosen ontologies, the context can be defined as any part of the ontologies that includes them. This means that the context can be either minimal, including mapped classes and the ones related to them, or extended in which case it includes more classes and consequently more information.

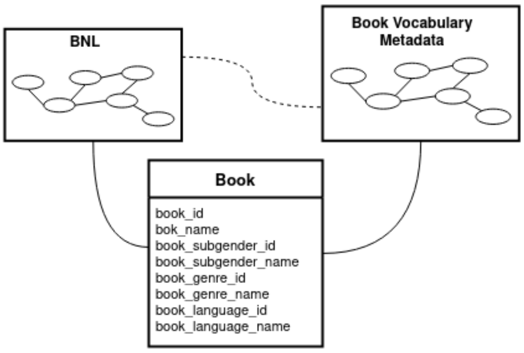


Figure 2: Book Dimension Mapping Example

Rules for DQ metrics are defined considering the DW model, the context and the mappings. A set of rules for syntactic accuracy for the category *name* of the Book dimension according to the property `dct:title` of the "British National Library" ontology is presented in equations 1 and 2, where in the predicate $SyntAcc(b, n)$, b is a particular book and $n \in \{0, 1\}$ is the result of the metric.

$$b \in Book\ Dimension \wedge b.name \in range(dct : title) \rightarrow SyntAcc(b, 1) \tag{1}$$

$$b \in Book\ Dimension \wedge b.name \notin range(dct : title) \rightarrow SyntAcc(b, 0) \tag{2}$$

The complete formalizations for the concepts presented above are implemented in Python using PyDatalog[3] for managing the DW model and DQ rules and owlready2[4] for managing ontologies.

4. Conclusions and Next Steps

The main strategy of our approach is based on the use and interaction between ontologies and Datalog, such that their reasoning power can be exploited for DQ rules. To the best of our knowledge, this approach has not been used before for this kind of solutions.

Up to now we completed first proposals of the literature review; a formalization for the DW based on [21] model; and a definition and formalization of the context based on domain ontologies. Following these first steps we proposed a mapping between the DW and the context and along with it, a way of managing the context scope, as a way of determining how much of the domain is being taken into account to give context to a DW. We worked on the implementation of each of the formalized

[1] http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf
[2] http://www.ebusiness-unibw.org/ontologies/opdm/book.html
[3] https://pypi.org/project/pyDatalog/
[4] https://pypi.org/project/Owlready2/

solutions. Finally, we defined and implemented a simple DQ metric for syntactic accuracy dimension in order to test the viability of the proposed solution.

The main focus of our ongoing work is to reach a level of abstraction in the formalization of the DW, the context and their interactions that makes it possible to evaluate certain DQ dimensions for any DW in any context. Currently, we are working on the definition and formalization of complex and generic DQ rules for consistency, completeness and accuracy.

Next steps will concentrate in the implementation of the complete solution based on the proposed formalizations, which will allow the definition of any DW, context and DQ rules set, as well as the application of the solution to a real case study.

References

[1] C. Batini, M. Scannapieco, Data and Information Quality, Data-Centric Systems and Applications, Springer International Publishing, Cham, 2016. URL: http://link.springer.com/10.1007/978-3-319-24106-7, dOI: 10.1007/978-3-319-24106-7.

[2] L. Bertossi, F. Rizzolo, L. Jiang, Data Quality Is Context Dependent, in: Enabling Real-Time Business Intelligence, Lecture Notes in Business Information Processing, Springer, Berlin, 2010, pp. 52–67. URL: https://link.springer.com/chapter/10.1007/978-3-642-22970-1_5. doi:10.1007/978-3-642-22970-1_5.

[3] M. Helfert, O. Foley, A Context Aware Information Quality Framework, in: 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009, pp. 187–193. doi:10.1109/COINFO.2009.65.

[4] A. L. McNab, D. A. Ladd, Information Quality: The Importance of Context and Trade-Offs, in: 2014 47th Hawaii International Conference on System Sciences, 2014, pp. 3525–3532. doi:10.1109/HICSS.2014.439.

[5] D. M. Strong, Y. W. Lee, R. Y. Wang, Data Quality in Context, Commun. ACM 40 (1997) 103–110. URL: http://doi.acm.org/10.1145/253769.253804. doi:10.1145/253769.253804.

[6] F. Serra, Handling Context in Data Quality Management, in: L. Bellatreche (Ed.), ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, Communications in Computer and Information Science, Springer International Publishing, Cham, 2020, pp. 362–367. doi:10.1007/978-3-030-55814-7_32.

[7] W. Fan, Data quality: From theory to practice, SIGMOD Rec. 44 (2015) 7–18. URL: https://doi.org/10.1145/2854006.2854008. doi:10.1145/2854006.2854008.

[8] P. Dourish, What we talk about when we talk about context, Personal and Ubiquitous Computing 8 (2004) 19–30. URL: https://link.springer.com/article/10.1007/s00779-003-0253-8. doi:10.1007/s00779-003-0253-8.

[9] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, P. Steggles, Towards a Better Understanding of Context and Context-Awareness, in: Handheld and Ubiquitous Computing, Lecture Notes in Computer Science, Springer, Berlin, 1999, pp. 304–307. URL: https://link.springer.com/chapter/10.1007/3-540-48157-5_29. doi:10.1007/3-540-48157-5_29.

[10] Y. W. Lee, Crafting Rules: Context-Reflective Data Quality Problem Solving, Journal of Management Information Systems 20 (2003) 93–119. URL: http://dx.doi.org/10.1080/07421222.2003.11045770. doi:10.1080/07421222.2003.11045770.

[11] X. H. Wang, D. Q. Zhang, T. Gu, H. K. Pung, Ontology based context modeling and reasoning using OWL, in: IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second, 2004, pp. 18–22. doi:10.1109/PERCOMW.2004.1276898.

[12] V. Luna, R. Quintero, M. Torres, M. Moreno-Ibarra, G. Guzmán, I. Escamilla, An ontology-based approach for representing the interaction process between user profile and its context for collaborative learning environments, Computers in Human Behavior 51 (2015) 1387–1394.

[13] C. A. Yeung, H. Leung, Formalizing typicality of objects and context-sensitivity in ontologies, in: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems, AAMAS '06, Assoc. for Computing Machinery, NY, USA, 2006, pp. 946–948. URL: https://doi.org/10.1145/1160633.1160801. doi:10.1145/1160633.1160801.

[14] N. Hernandez, J. Mothe, C. Chrisment, D. Egret, Modeling context through domain ontologies, Information Retrieval 10 (2007) 143–172. URL: https://doi.org/10.1007/s10791-006-9018-0. doi:10.1007/s10791-006-9018-0.

[15] O. Barkat, S. Khouri, L. Bellatreche, N. Boustia, Bridging context and data warehouses through ontologies, in: Proceedings of the Symposium on Applied Computing, SAC '17, Association for Computing Machinery, NY, USA, 2017, p. 336–341. URL: https://doi.org/10.1145/3019612.3019838. doi:10.1145/3019612.3019838.

[16] A. Ranganathan, R. H. Campbell, An infrastructure for context-awareness based on first order logic, Personal and Ubiquitous Computing 7 (2003)

353–364.

[17] F. Serra, A. Marotta, Context-based Data Quality Metrics in Data Warehouse Systems, CLEI Electronic Journal 20 (2017) 3:1–3:23. URL: https://www.clei.org/cleiej/index.php/cleiej/article/view/22. doi:10.19153/cleiej.20.2.3, number: 2.

[18] S. Abdellaoui, L. Bellatreche, F. Nader, A quality-driven approach for building heterogeneous distributed databases: The case of data warehouses, in: 2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2016, pp. 631–638. doi:10.1109/CCGrid.2016.79.

[19] L. Bertossi, M. Milani, Ontological Multidimensional Data Models and Contextual Data Quality, J. Data and Information Quality 9 (2018) 14:1–14:36.

[20] A. Marotta, A. Vaisman, Rule-Based Multidimensional Data Quality Assessment Using Contexts, in: Big Data Analytics and Knowledge Discovery, Lecture Notes in Computer Science, Springer, Cham, 2016, pp. 299–313.

[21] C. A. Hurtado, A. O. Mendelzon, OLAP Dimension Constraints, in: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM, NY, USA, 2002, pp. 169–179.

[22] C. A. Hurtado, A. O. Mendelzon, Reasoning about Summarizability in Heterogeneous Multidimensional Schemas, in: J. Van den Bussche, V. Vianu (Eds.), Database Theory — ICDT 2001, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2001, pp. 375–389. doi:10.1007/3-540-44503-X_24.