Towards a Knowledge Graph to Describe and Process Data Defects

João Marcelo Borovina Josko*, Lisa Ehrlinger†‡, Wolfram Wöß†
*Federal University of ABC, Av. dos Estados, 5001 Bairro Santa Terezinha – Santo André, Brazil
email: marcelo.josko@ufabc.edu.br
†Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria
email: lisa.ehrlinger@jku.at, wolfram.woess@jku.at
‡Software Competence Center Hagenberg, Softwarepark 21, 4232 Hagenberg, Austria
email: lisa.ehrlinger@scch.at

***Abstract*—The reliability and trustworthiness of machine learning models depends directly on the data used to train them. Knowledge about data defects that affect machine learning models is most often considered implicitly by data analysts, but usually no centralized data defect management exists. Knowledge graphs are a powerful tool to capture, structure, evolve, and share semantics about data defects. In this paper, we present an ontology to describe data defects and demonstrate its applicability to build a large public or enterprise knowledge graph.**

Keyword Terms— Data Defects; Data Quality Assessment; Knowledge Graphs.

I. Introduction

If the data used for Machine Learning (ML) applications has defects, the resulting ML model will perform poorly and generate unreliable results. Possible effects are cost increase, incorrect decision making, customer dissatisfaction, and organizational mistrust within organizations [1]. Examples for data defects, which have received increased attention in the ML community, are missing data (by error) and outlying values [2]. However, knowledge about such defects is almost always tacit within organizations and concentrated on a few data professionals that may have an incomplete understanding of all data defect implications and characteristics. Knowledge Graphs (KGs) bear the potential to capture, structure, evolve, and share semantics about data defects, which constitutes the basis for comprehensive Data Quality (DQ) management for ML applications. DQ is most often associated with the "fitness for use" principle [3][4], which highlights the importance of taking into account the respective context and the consumer (i.e., user or service) of the data. While there has been a lot discussion on Data Quality Assessment (DQA) in general (cf. [1][5][6]), and data defects in particular (cf. [7]–[9]), an analysis of the literature reveals a segmented representation of data defect knowledge.

KGs, which are defined to "acquire and integrate information into an ontology and apply a reasoner to derive new knowledge" [10], have already been successfully applied to organize the semantic information of different domains, like scientific documents [11][12]. However, so far, there exists no KG to describe data defects. To address this gap, we present a KG model in form of an ontology to represent the semantic information of data defects and show how to apply it to public or enterprise KGs, i.e., how to populate such a KG. The main contribution of this paper is an ontology that allows a practitioner to know *which* knowledge about data defects is required and *how* to organize it. The high expressiveness of ontologies [13] allows to incorporate the context (cf. fitness-for-use principle of DQ [3]) of the data defects within the model, such as the function or database (DB) table where a defect occurs.

This paper is structured in three parts: Section II provides an overview on related work. Section III comprises a theoretical introduction to data defects, the discussion of our ontology, and its applicability. We conclude in Section IV.

II. Related Work

DQ literature provides huge knowledge about data defects, with certain papers discussing topics like data defect structures [7][8][14], methods of data defect detection [6], DQ dimensions [1][6][3] and DQA process characterization [1][5][9]. Despite their considerable contribution, no attention has been paid to represent the relationships among data defect concepts and the situation they appear in (i.e., context). In ML applications, explicit knowledge about data defects, like missing or outlying values, would enhance prediction accuracy. To incorporate knowledge about data defects into ML models, it is thus necessary, to describe it semantically.

KGs have already been successfully applied to describe complex domains like science [11][12] or the Italien cultural heritage within the ArCo project [15]. Following this line, some works provide a semantic description of the DQ assessment domain, observing the topic from a general [16] or domain-specific [17] perspective (e.g., linked data). However, these works focus on the task of assessing or measuring DQ and do not go into detail to describe specific data defects. In this paper, we provide a machine-readable semantic representation of the data defect domain, which provides on the one hand a standardized and centralized repository about data defects and their handling, and on the other hand, allows to incorporate this knowledge into automated ML workflows.

III. A Knowledge Graph to Describe Data Defects

In this section, we (1) explain the theory behind data defects, (2) present an ontology on data defects, which constitutes the structure of a KG, and (3) demonstrate how to apply this ontology and build a public or enterprise KG.

 ISBN: 978-1-61208-715-3

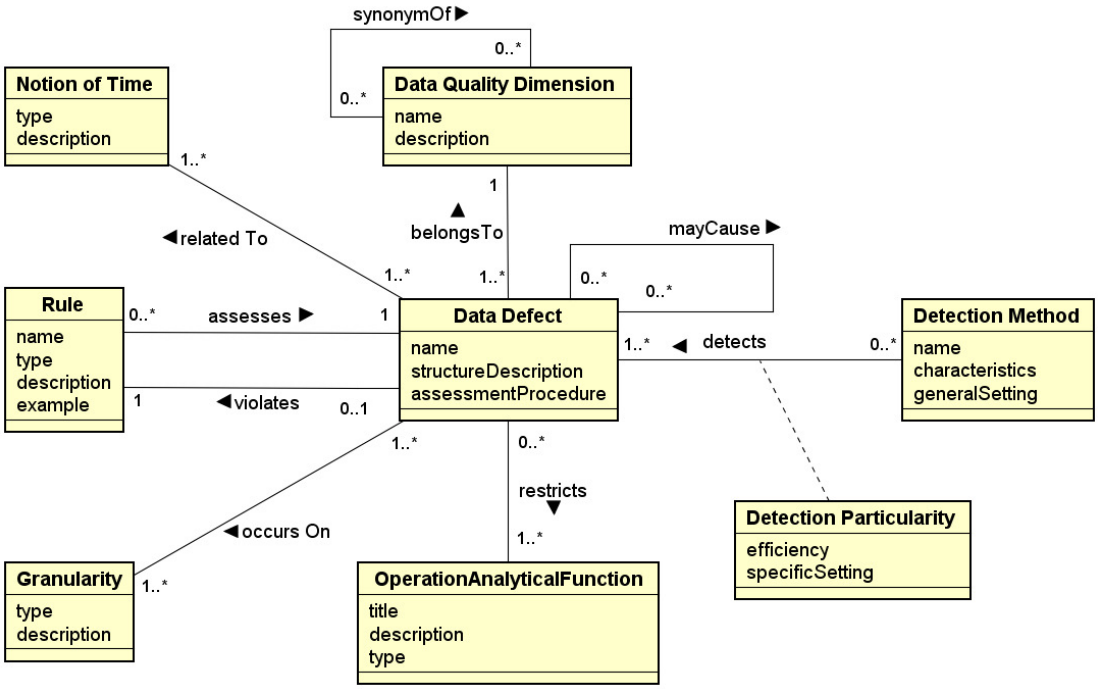

Fig. 1. An ontology to represent the data defect domain

A. Theory on Data Defects

In a nutshell, a data defect is a disagreement between what is provided by a database and what is expected from it according to some data semantic. Such disagreement results from rule violations like organizational business rules (e.g., domain rule, tax rules) or database implicit rules (e.g., databases should not have duplicates) [9]. The way in which a rule is violated denotes the structure of a particular data defect [8].

Data defects also share some implicit properties. The first property refers to *inherently complex nature*. A data defect occurs in more than one granularity (e.g., value, tuple, column, relation), and its core structure may possess slight variations or particularities. Moreover, in certain data settings, a data defect $\mathcal{D}_{\mathcal{A}}$ may cause a defect $\mathcal{D}_{\mathcal{B}}$ and, progressively, lead to a chain reaction [5]. This situation can be especially critical in the case of temporal data.

The next property refers to *level of human supervision* required to determine a data defect [5]. While some defects can be automatically determined through assessment rules (as used by data profiling tools), other data defects require knowledge about a particular business context to be refuted or confirmed. In any case, each data defect demands a particular assessment analysis procedure.

As the last property, data defects can also cause distinct *impacts on data life cycle operations* (e.g., use, maintain or purge data) and, consequently, operational and analytical functions they are part of. Certain defects may totally obstruct one or more functionalities such as credit concession blockage for certain customer on account of an incorrect income value. In contrast, other "less severe" data defects do not inhibit functionalities, but they "use" them to proliferate defective data all around organizational databases. An example of this case refers to determine product discounts based on incorrect customers ranking.

B. The Data Defect Ontology

Figure 1 shows the ontology (diagrammed in UML) that provides concepts and constructs for specifying, organizing, evolving, and communicating semantic content about data defects, according to data defects properties (Section III-A). Its key concept is *Data Defect*. It represents a violation of a *Rule* that leads to defective data. Conversely, a rule (or set of rules) may be used to discover a data defect. Moreover, a data defect belongs to a particular *Data Quality Dimension* (e.g., accuracy, consistency, as proposed in [3]) and refers to some *Notion (Dimension) of Time* like snapshot, valid time and transactional time [7].

The connection to the data is provided by the concept of *Granularity*, which defines a specific granularity of the data, where a data defect can occur on (e.g., value, tuple, column, or relation in a database). This granularity can for example be specified with a SQL statement that links to the affected data.

The presence of defective data has the potential to restrict the use of a number of *Operational and Analytical Functions* (OperationalAnalyticalFunction). Besides these impacts, the ontology also models impacts between data defects, i.e., the fact that certain data defects may trigger other data defects. The current version does not contain the impact of the user or service using the data, which is part of our future work.

 ISBN: 978-1-61208-715-3

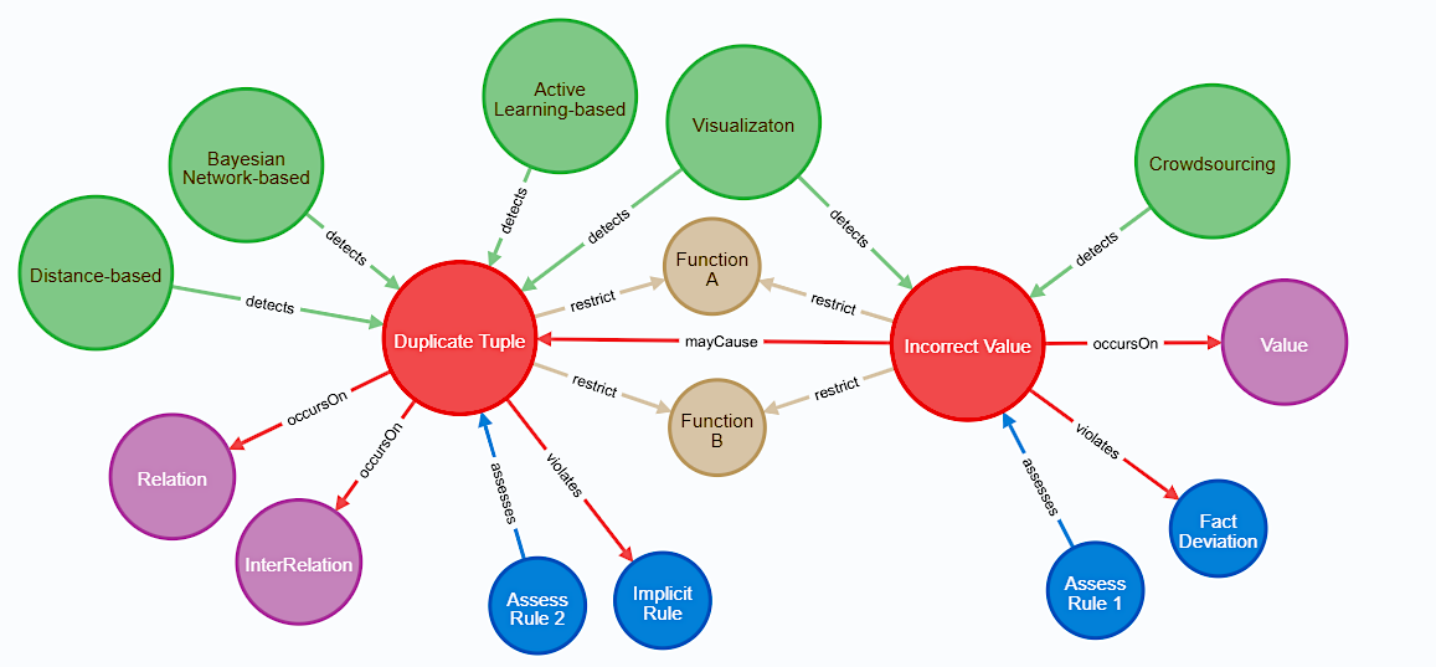

Fig. 2. A KG prototype for the data defect domain

A *Detection Method* can reveal a specific set of data defects, but with different efficiency and configuration setting (*Detection Particularity*). However, each data defect has a particular human assessment analysis procedure (*assessmentProcedure*). Further, a data defect can occur on one ore more granularity levels, such as values, tuples, columns, or relations. On the one hand, defects can be assessed through rules (*Rule*), which are, e.g., used by data profiling tools. On the other hand, data defects can be characterized by the rules they violate. This interrelation is modeled by the two relationships between rules and data defects.

Our model does not intend to represent semantically expected data volatility situations (e.g., missing data values that do not exist in the real world) since they do not represent a data defect (cf. Section III-A). In addition we want to point out that the focus of our research is on DQ assessment (detection, measurement) and automatic data cleansing activities are not in the scope of this research work. However, it is necessary to measure and know the quality of the data to understand the degree and effectiveness of data cleansing and to define goals for further cleansing activities [18][19]. The incorporation of data cleansing is planned in future work, but it requires a deeper investigation of data cleansing methods to expand the model appropriately.

C. Application with a Knowledge Graph Prototype

To demonstrate the applicability of our data defect ontology, we built a KG prototype using the Neo4J graph database [20]. Figure 2 shows an exemplary query result that highlights the consistency and clarity provided by the data defect ontology. In order to keep Figure 2 readable, the query has been restricted to a subset of properties and concepts about two notorious data defects: *Duplicate Tuples* and *Incorrect Values*. Further details about these and other timeless data defects are discussed in [8].

Each node color in Figure 2 corresponds to one concept expressed in the data defect ontology (cf. Section III-B) and each edge color corresponds to its source node color and the label exhibits its role. We used the following color code:

- Red: data defect
- Blue: rule
- Purple: granularity
- Brown: operational-analytical function
- Green: detection method

The nomenclature used for the labels of the nodes for data defects (red), granularity (purple), and detection method (green), exhibit name-based attributes that are notorious in database literature like [2][8]. Further information on the two data defects *Duplicate Tuples* and *Incorrect Values* is provided in [8]. A *Relation* refers to a table in a relational database, *InterRelation* to a join between two or more tables, and a *Value* to one specific value within a table (e.g., a integer, string, or boolean). While general information on different data defect detection methods is reviewed by Dasu and Johnson [9], methods specifically attributed to duplicate detection are summarized by Elmagarmid et al. [21]. A few examples are *Distance-based* methods, which are based on a function that calculates the distance between two objects [21], *Bayesian-Network-based* methods, which compute the conditional dependencies between objects (variables) using probabilistic inference [22], *Active-Learning-based* methods that rely on ML methods, *Visualization* as important tool for detection data defects (as highlighted in [23]), and *Crowdsourcing*.

 ISBN: 978-1-61208-715-3

Since business rules and operational-analytical functions (cf. Section III-A) rely on the domain context, their corresponding nodes in Figure 2 (blue and brown respectively) use fictional labels, e.g., "Assess Rule 1", or "Function A". To maintain the figure readable and demonstrative, we did not include *Notion of Time* in the current version.

IV. Conclusion and Future Work

This paper introduces an ontology to represent semantic knowledge about data defects, which extends the W3C DQ vocabulary [16]. The design of the ontology considers several data defect properties and its applicability was examined by means of a KG prototype. Such a knowledge graph enables organizations to acquire, organize, evolve, and promote a common understanding of data defects within their domain. In future works, we intend to (1) investigate how knowledge regarding the data defects domain can be captured automatically, (2) additionally take into account the impact on DQ from the user or service utilizing the data, and (3) extend our ontology to fully support spatial data defects. The latter refers to the ability to full express data defects semantics with respect to relationships among distinct spatial attributes (e.g., location, shape, size, and orientation) that are not properly captured by *Granularity*, for instance.

Acknowledgment

The research reported in this paper has been partly supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Digital and Economic Affairs, and the Province of Upper Austria in the frame of the COMET center SCCH.

References

[1] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Communications of the ACM*, vol. 41, no. 2, 1998, pp. 79–82.

[2] C. C. Aggarwal, *Outlier analysis*. Springer International Publishing, 2017.

[3] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, March 1996, pp. 5–33.

[4] N. R. Chrisman, "The role of quality information in the long-term functioning of a geographic information system," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 21, no. 2, 1983, pp. 79–88.

[5] J. M. Borovina Josko, "Uso de propriedades visuais-interativas na avaliação da qualidade de dados (in portuguese)," Ph.D. thesis, Universidade de São Paulo, 2016.

[6] D. Loshin, *The practitioner's guide to data quality improvement*. Elsevier, 2010.

[7] J. M. Borovina Josko, "A formal taxonomy of temporal data defects," in *Data Quality and Trust in Big Data*, H. Hacid, Q. Z. Sheng, T. Yoshida, A. Sarkheyli, and R. Zhou, Eds. Cham: Springer International Publishing, 2019, pp. 94–110.

[8] J. M. Borovina Josko, M. K. Oikawa, and J. E. Ferreira, "A formal taxonomy to improve data defect description," in *International Conference on Database Systems for Advanced Applications*. Springer, 2016, pp. 307–320.

[9] T. Dasu, "Data glitches: Monsters in your data," in *Handbook of Data Quality*. Heidelberg, Germany: Springer, 2013, pp. 163–178.

[10] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs," in *Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems - SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS16)*, ser. CEUR Workshop Proceedings, E. F. Michael Martin, Martí Cuquet, Ed., vol. 1695, Aachen, 2016, pp. 13–16.

[11] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, and M. E. Vidal, "Towards a knowledge graph for science," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS '18)*. New York, NY, USA: ACM, 2018, pp. 1:1–1:6.

[12] S. Fathalla, S. Vahdati, S. Auer, and C. Lange, "Towards a knowledge graph representing research findings by semantifying survey articles," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2017, pp. 315–327.

[13] C. Feilmayr and W. Wöß, "An analysis of ontologies and their success factors for application to business," *Data & Knowledge Engineering*, vol. 101, 2016, pp. 1–23.

[14] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Engineering Bulletin*, vol. 23, no. 4, 2000, pp. 3–13.

[15] L'Istituto Centrale per il catalogo e la Documentazione (ICCD), "ArCo," 2019, https://github.com/ICCD-MiBACT/ArCo [retrieved: May, 2019].

[16] R. Albertoni and A. Isaac, "Data on the web best practices: Data quality vocabulary," *W3C Working Draft*, vol. 19, 2016.

[17] J. Debattista, C. Lange, and S. Auer, "daq, an ontology for dataset quality information," in *Linked Data on the Web (LDOW)*, 2014.

[18] L. Sebastian-Coleman, *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Newnes, 2012.

[19] L. Ehrlinger, B. Werth, and W. Wöß, "Automated Continuous Data Quality Measurement with QuaIIe," *International Journal on Advances in Software*, vol. 11, no. 3 & 4, 2018, pp. 400–417.

[20] J. J. Miller, "Graph database applications and concepts with Neo4j," in *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, vol. 2324, 2013.

[21] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, 2007, pp. 1–16.

[22] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[23] C. Bors, T. Gschwandtner, S. Kriglstein, S. Miksch, and M. Pohl, "Visual interactive creation, customization, and analysis of data quality metrics," *Journal of Data and Information Quality (JDIQ)*, vol. 10, no. 1, 2018, p. 3.

 ISBN: 978-1-61208-715-3