11th International Conference on Through-life Engineering Services – TESConf2022
8-9 November 2022, Cranfield UK

Developing an Ontological Framework for Effective Data Quality Assessment and Knowledge Modelling

Christina Latsou[a]*, Marta Garcia I Minguell[a], Ayse Nur Sonmez[a], Roger Orteu I Irurre[a], Martin Mark Palmisano[a], Suresh Landon-Valdez[a], John Ahmet Erkoyuncu[a], Pavan Addepalli[a], Jim Sibson[b], Olly Silvey[b]

[a]*Centre for Digital Engineering and Manufacturing, College Rd, Cranfield University, MK43 0AL, UK*
[b]*Babcock International Group Plc, Ashton Vale Rd, Bristol BS3 2HQ, UK*

* Corresponding author. Tel.: +44-1234-750-111. *E-mail address:* christina.latsou@cranfield.ac.uk

Abstract

Big data has become a major challenge in the 21st century, with research being carried out to classify, mine and extract knowledge from data obtained from disparate sources. Abundant data sources with non-standard structures complicate even more the arduous process of data integration. Currently, the major requirement is to understand the data available and detect data quality issues, with research being conducted to establish data quality assessment methods. Further, the focus is to improve data quality and maturity so that early onset of problems can be predicted and handled effectively. However, the literature highlights that comprehensive analysis, and research of data quality standards and assessment methods are still lacking. To handle these challenges, this paper presents a structured framework to standardise the process of assessing the quality of data and modelling the knowledge obtained from such an assessment by implementing an ontology. The main steps of the framework are: (i) identify user's requirements; (ii) measure the quality of data considering data quality issues, dimensions and their metrics, and visualise this information into a data quality assessment (DQA) report; and (iii) capture the knowledge from the DQA report using an ontology that models the DQA insights in a standard reusable way. Following the proposed framework, an Excel-based tool to measure the quality of data and identify emerging issues is developed. An ontology, created in Protégé, provides a standard structure to model the data quality insights obtained from the assessment, while it is frequently updated to enrich captured knowledge, reducing time and costs for future projects. An industrial case study in the context of Through life Engineering Services, using operational data of high value engineering assets, is employed to validate the proposed ontological framework and tool; the results show a well-structured guide that can effectively assess data quality and model knowledge.

Keywords: data quality issues; data quality dimensions; data quality assessment; ontology; data management

1. Introduction

Big data has currently become a major challenge, demanding advanced and cost-effective methods of information processing to achieve enhanced insights, informed decision-making, and process automation [1]. The current literature gives great emphasis on the classification, mining and extraction of knowledge from data obtained from disparate sources. To increase the confidence in data-driven decision making, measuring, assessing and improving the quality of the data is fundamental [1][2]. Data quality assessment (DQA) is, thus, vital for organisations to understand the customer needs effectively, gain enhanced insights, improve services, and predict and prevent risks [3]. However, it is estimated that data analysts devote on average 40% of their time solving data quality (DQ) issues while low DQ costs companies £12 million per year on average [4].

Considering the existing literature, large amount of data from abundant data sources with non-standard structures and limited timeliness makes the development of methodological approaches for assessing the quality of data and establishing DQ standards a challenging task [5]. With this regard, establishing assessment methods able to detect DQ issues, determine the level of DQ, evaluate the usefulness of data to users and improve the quality of data has become crucial. Additionally, ontology-driven approaches could advance this DQA journey to represent shareable and reusable knowledge across a domain. However, a lack of studies proposing a comprehensive methodological approach able to assess the quality of data regardless of the context of usage has been identified in the literature [1].

To handle this challenge, this work addresses the following research question: "How to develop a framework to structure the process of assessing the quality of data and modelling the knowledge obtained from the assessment?". This study contributes to knowledge by: (i) proposing a methodological framework to assess the quality of

Peer-review under responsibility of the Programme Chair of the 11th International Conference on Through-life Engineering Services. © Licence Under: CC-BY-4.0

data, while enabling generic applicability; and (ii) developing an ontology to capture the knowledge obtained from the DQA. The proposed ontology is reusable once source data is updated and provides a common understanding of the structure of insights obtained from the DQA.

The remainder of the paper is structured as: Section 2 discusses the literature review, Section 3 presents the proposed DQA framework, Section 4 validates the framework through a case study, and Section 5 provides a conclusion to the paper.

2. Literature review

Research on DQ has been conducted since the 1980s and is being associated to the ‘fitness for use’ principle, i.e., the quality of data is defined by the user according to their requirements and type of usage [6]. DQ is understood as a multi-dimensional concept where *dimensions* and *metrics* play key role. *Dimensions* are categories of DQ issues with a shared reason why they are important and may have similar underlying causes (i.e. what to measure). *Metrics* describe how a dimension is quantified, quantitatively and/or qualitatively (i.e. how to measure dimensions) [7]. Hence, DQ dimensions are assessed via metrics. Dimensions, classifications and metrics are extensively discussed in the literature. However, there is no consensus on what makes a good dataset of DQ dimensions and a standard definition for each dimension [7-9].

2.1 Data quality assessment

DQA, a dynamic process that changes once source data is updated [10], extends the concept of DQ measurement by appraising the results of measurement to obtain insights and make decisions about the object of assessment [11]. Over the years, a number of research studies discuss the topic of DQA across a wide spectrum of contexts in the literature. Research studies contributed to the topic of DQA are further discussed. Pipino et al. [12] proposed a novel methodology by introducing the concept of subjective and objective assessments of DQ. Subjective assessment reflects the needs and experiences of users, whereas objective covers the states of the data without the contextual knowledge of the application, business rules and constraints. This study proposes a structured approach where the subjective and objective assessments are compared to identify discrepancies and improve the quality of data. Moreover, Bergdahl et al. [13] developed handbook to facilitate a systematic implementation of DQA in the European Statistical System. This work discusses assessment methods, considering data quality reports, process variables and indicators. Methods including audits, self-assessment and peer reviews that rely on information from quality indicators, reports, process variables and user surveys are also considered. Furthermore, Batini et al. [14] conducted a comprehensive comparison of several DQ classifications considering thirteen methodologies. They concluded that there is no agreement either on which set of dimensions defines the quality of data, or on the exact meaning of each dimension. In more recent DQA literature, Camera et al. [15] introduced a service DQ framework to provide companies with a set of methods and tools to prioritise relevant service data and assess its quality levels.

Cichy and Rass [5] provided a comprehensive overview on applicable DQ methodologies, while Ehrlinger and Wöß [1] conducted a systematic survey on the topic of DQ, concluding that the methodologies in the current literature employ abstract dimensions with no common understanding. Ehrlinger and Wöß [1] also highlighted the need for more automation in DQ measurement and comprehensive explanation of the calculations and algorithms. To conclude, research on DQ has provided numerous approaches to guide organisations in the assessment, analysis and improvement of DQ dimensions. However, there is still no consensus on a standardised list of dimensions and metrics for DQA and the topic of DQ is closely related to the elements of subjectivity and context-dependency [1][16].

2.2 Data quality ontology

Developing an ontology to capture insights from DQ measurement or assessment is a relatively new concept in the literature. Research on ontology and DQ are, so far, studied in isolation. In 2014, Debattista et al. [17] proposed the dataset quality ontology (daQ), a vocabulary with metrics for measuring the quality of a dataset. However, this work represents only the early stages of developing a DQ ontology, requiring comprehensive and further modelling to be capable of representing the domain of interest. Moreover, in 2015, Johnson et al. [18] proposed an ontology to define DQ dimensions in healthcare. The aforementioned ontologies help more with publishing quality reports in a machine-readable manner, rather assessing the quality of data though. More recently, in 2022, Nayak et al. [19] developed a preliminary ontological framework to build an end-to-end system able to assess and improve the DQ by identifying the root causes of DQ violations.

2.3 Research gap

Overall, the existing literature on DQA is sparse and limited to individual methods. Moreover, the existing findings for DQA methods are based on the context of data usage, lacking generic applicability. Thus, there has been identified a lack of research evidence on approaches that methodologically assess the quality of data and formally capture the knowledge obtained from the assessment. In terms of this latter point, limited research has been conducted on using the concept of ontology to model the knowledge obtained from the quality assessment of a dataset. Ontology, an object-oriented approach, could help share information in a domain, while reuse and enrich captured knowledge. Therefore, this research work aims to present a structured framework for assessing the quality of data and formally model the knowledge from the assessment regardless of the context of data usability.

3. Data quality assessment framework

In this section, the steps followed for the development of the proposed design framework for DQA and formal knowledge modelling are discussed. The framework consists of three steps: Step 1 captures user's requirements identifying what data and how it should be assessed; Step 2 assesses the level of data quality and visualises the insights obtained from the assessment; and Step 3 models the DQA insights to a data quality ontology. An abstract view of these steps is illustrated in Fig.1. The framework develops a structured guide that can standardise the process of assessing the quality of data and modelling the knowledge obtained from such an assessment by implementing an ontology.

3.1. Step 1: User's Requirements Identification

The aim of this step is to understand the context of a dataset and how the data is to be used. Users' requirements in terms of data values, format and ranges are identified. The output of Step 1 is two forms, the context form and data requirement form, that will be used in Step 2 to assess the DQ by comparing the users' expectations against current state of data. Two sub-steps are considered, as:

Step 1.1 - Context identification form: helps users understand the DQA scope and what needs to be measured. A set of questions is considered to identify the aim, goals and data needs to be assessed.

Step 1.2 - Data requirements identification form: captures the users' requirements, considering data types, formats and thresholds, for assessing the DQ. These requirements help define rules, constraints and relationships for the data. Two types of requirements are considered: (i) schema requirements, applied to the dataset schema; and (ii) column requirements, applied to the content (values) of dataset's columns, as defined by the user.

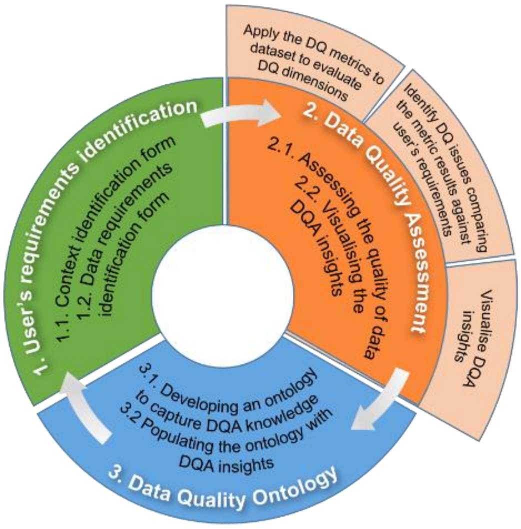
1. User's requirements identification
1.1. Context identification form
1.2. Data requirements identification form
2. Data Quality Assessment
2.1. Assessing the quality of data
2.2. Visualising the DQA insights
Apply the DQ metrics to dataset to evaluate DQ dimensions
Identify DQ issues comparing the metric results against user's requirements
Visualise DQA insights
3. Data Quality Ontology
3.1 Developing an ontology to capture DQA knowledge
3.2 Populating the ontology with DQA insights

Fig. 1. Data quality assessment framework.

3.2. Step 2: Data Quality Assessment

After capturing the user's requirements in Step 1, this step focuses on assessing the level of DQ. The output of this step is a report that presents the insights obtained from the assessment of the quality of a given dataset. DQA is formulated as:

Step 2.1 – Assessing the quality of data: The aim of this step is to measure and evaluate the level of DQ of a dataset considering the users' inputs as captured in the data requirement form (Step 1.2) to support decision-making. In this step, dimensions, including timeliness, currency, accuracy, completeness, consistency, uniqueness and validity, are evaluated. These dimensions have been selected as being the most cited in the literature, demonstrating their importance [7][15]. To evaluate each dimension a metric has been developed and applied to the dataset. Table 1 shows the definitions and metrics for each data quality dimension. Qualitative dimensions, such as timeliness and currency, are assessed based on the users' requirements from Step 1.2, whereas quantitative dimensions are calculated by dividing the results of metric to the total number of rows. Metrics are also measured as percentages and then compared against the user's requirements to evaluate the level of DQ.

Table 1. Data quality dimensions, definitions and metrics.

Dimension	Definition	Metric
Timeliness	The period between when information is expected and readily available for use.	Difference between actual and scheduled time of data delivery.
Currency	The degree to which data is up to date.	Difference between the real state and state data captured.
Accuracy	The degree to which data mirrors the characteristics of the	Number of data entries that pass the

	real world (thresholds).	data accuracy rules or ranges.
Completeness	The absence of blank values.	Number of rows not in blank.
Consistency	The absence of difference by assessing pattern.	Number of rows with equal standardised text.
Uniqueness	The unwanted duplication for a particular record.	Number of unique rows.
Validity	The degree to which the data conforms to the data types and format.	Number of valid rows (i.e. string, integer, etc.).

Step 2.2 – Visualising the DQA insights: The results obtained from Step 2.1 are presented in a DQA report, demonstrating the state of quality of the dataset using tables and charts. This communicates the DQ insights to users and supports informed decision-making. The DQA can reveal issues that can be further investigated to identify and address the root causes, and prevent future losses.

Step 3: Data Quality Ontology Development

In this framework, an ontology provides a standard structure to model the DQA and insights obtained from Step 2, while it can be updated to enrich captured knowledge. Step 3 is carried out as:

Step 3.1 - Developing an ontology to capture DQA knowledge: The design of the ontology starts with a DQA process to measure and evaluate the quality of a dataset. This process that represents the highest level of DQA, contains two types of DQ classifications, qualitative and quantitative, based on the metric used to evaluate each dimension. Each classification has a definition and contains dimension(s). Each dimension has a dimensional definition and is assessed through metric(s). A DQ metric has a definition, data type (e.g. integer, Boolean, etc.) and a value obtained from the assessment, while contains metric units and an insight type of data state that indicates the condition of the data after is assessed and compared against the requirements captured in Step 1. The state is expressed as 'accepted' or not accepted. A UML class diagram of the proposed DQ ontology is demonstrated in Fig. 2.

Step 3.2 - Populating the ontology with DQA insights: the DQA insights obtained from Step 2 are populated into the ontology model as discussed in Step 3.1. Once source data, in Step 1, is updated and new insights are obtained after the assessment in Step 2, these can be modelled into the ontology updating the previously captured knowledge. This cyclic process, seen in Fig. 1, can reduce time and cost related to DQA.

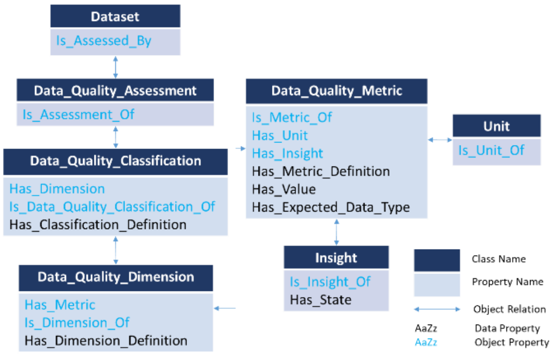

Fig. 2. UML class diagram of the data quality ontology.

4. Data quality assessment toolkit

Following the proposed framework, a step-by-step dynamic Excel-based toolkit to assess the quality of data and identify emerging issues is developed. An ontology, developed in Protégé, provides a standard structure to model the DQ insights obtained from the assessment. A case study in the context of Through life Engineering Services, using operational data of high value engineering assets, is employed to validate the proposed ontological framework. This section discusses the development of the toolkit, based on the three steps introduced in Section 3 for the design framework.

4.1. Step 1: User's Requirements Identification

To capture the user's requirements and gain understanding of the data context, two hourly interviews were conducted with the client's data analytics team. According to the proposed framework, the two sub-steps are, as:

Step 1.1 - Context identification form: The dataset provided relates to the management and development of a number of critical assets involved in global operations. It holds information for thirty four assets and five model types that operate in various locations. The dataset consists of 23 columns and 39569 rows. It contains information for the usage of assets over time, recoding data in terms of land and nautical mileage, average speed of land, and flight hours, average altitude, and minimum and maximum water temperatures. The challenge arises with the 'asset usage' dataset is that the data collected over the years has been merged into one file to have a single point of reference. This can affect the quality of data due to its heterogeneity, including structural and lexical differences across the merged datasets. The aim of the DQA is to assess the current DQ level of the assets and identify issues.

Step 1.2 - Data requirements identification form: after understanding the dataset context and DQA scope, the requirements for the dataset schema and columns, summarised in Table 2, have been identified, based on the customer's expectations. The 'schema requirements' are applied to the dataset schema and related to the dimensions of timeliness, completeness, uniqueness and validity. For instance, the requirements related to completeness and uniqueness express that no empty fields and duplicate values are expected in the schema dataset, respectively. Moreover, for each of the selected dimensions, the 'column requirements' are applied to the values of either the whole dataset or certain columns as defined by the user, as seen in Table 2. The requirement related to timeliness, for instance, indicates that all the data in all the columns should be evaluated, assessing if they are available for use, while the requirement related to currency shows that the values in 'Date' column should be 2022, proving the data is updated. For the accuracy of data, maximum thresholds for the values in 'maximum water temperature' column equal to 85ºC have been provided. It is expected that at least 90% of these values to be within the threshold. Moreover, for 'model type' column, the short standardised text input options 'M1', 'M2' and 'M3' have been captured. The values in this column should be in text format and equal to one of the three input options.

Table 2. Data requirements identification form

Dimension	Schema Requirement	Column in dataset	Column Requirement
Timeliness	Available	All	Available
Currency	-	Date	Year - 2022
Accuracy	-	Water temperature	≥ 90% within thresholds
Completeness	100% no blank	All	≥ 95% no blank
Consistency	-	Asset ID, model type	100% standard input
Uniqueness	100% of no duplicates	ID, Date	100% of no duplicates
Validity	100% string	All	100% with equal format

4.2 Step 2: Data Quality Assessment

In this step, the quality of 'asset usage' dataset is assessed based on the user's requirements identified in Step 1. According to the framework:

Step 2.1 - Assessing the quality of data: The quality of the dataset has been measured using the metrics from Table 1 for both the schema and column data. The schema data has been assessed successfully by confirming that it follows the schema requirements from Table 2. In terms of the column data, for each dimension, the DQ level, DQA results, obtained by comparing the results from the metrics with the column requirements, and states (i.e. accepted or not accepted) have been found, as viewed in Table 3. It can be seen that the accuracy and consistency of the dataset cannot be accepted, as they do not satisfy the required thresholds. The accuracy is below the threshold by 27.74% due to high values in the water temperatures, whereas the consistency is lower than the required threshold by 3.29% due to errors in the standardised text caused by manual data entries.

Table 3. DQA results based on the column requirements.

Dimension	DQ Level	Result	State
Timeliness	Available	Yes	Accepted
Currency	2035	Yes	Accepted
Accuracy	28592	72.26%	Not accepted
Completeness	38876	98.25%	Accepted
Consistency	38267	96.71%	Not accepted
Uniqueness	39569	100%	Accepted
Validity	39569	100%	Accepted

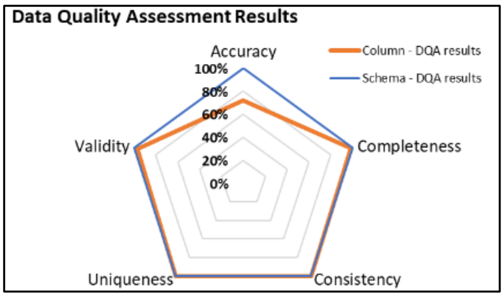

Fig. 3. Schema and column results obtained from DQA.

Step 2.2 - Visualising the DQA insights: The radar chart in Fig. 3 shows the results obtained from the schema and column DQA results in Step 2.1. According to the findings, the data accuracy should be examined further to identify the root causes. From the analysis, it was found that the water temperature was higher than the expected by 11ºC on average. This could have been caused due to high weather temperatures during the summer months or random failures during the assets' life.

4.3 Step 3: Data Quality Ontology Development

In this step, the DQ ontology is implemented as:

Step 3.1 - Developing an ontology to capture DQA knowledge: following the ontological design architecture presented in Fig. 2, an ontology for modelling the DQA knowledge is developed. The Protégé interface for DQ ontology is viewed in Fig. 4. The data structure of the DQA has been formalised with attributes and relationships to describe the DQA based on the quantitative and qualitative classifications, dimensions, metrics and insights. A 'dataset' class with a sub-class of 'asset usage' are also added for modelling the dataset of the selected case study.

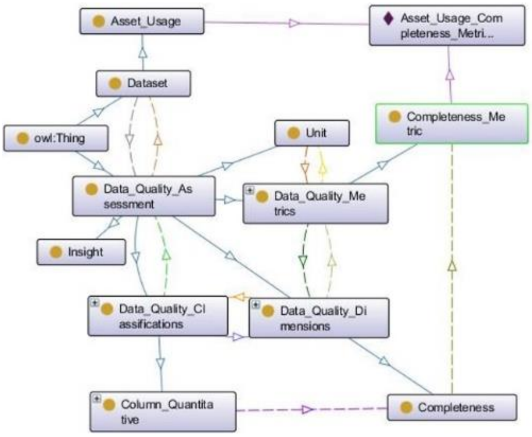

Fig. 4. Protégé interface for data quality ontology.

Step 3.2 - Populating the ontology with DQA insights: the DQA results and states for each dimension, obtained in Step 2.1, are stored to the DQ ontology using Owlready2. Moreover, ontologies have reasoning capabilities that allow for semantic queries to the stored data and, thus, inference of useful knowledge. This mechanism enables the retrieval of explicit and implicit knowledge derived from the semantic information associated with the data. For instance, a contextual query can be expressed as: “Which dimensions are not accepted based on the DQA insights?”. This automated process querying can help identify promptly DQ issues and support strategic decision-making.

5. Conclusions and future work

In this study, a structured framework for assessing DQ and help users understand and verify data usability is proposed. A DQ ontology with a design architecture that models the insights obtained from DQA is also proposed. A dynamic Excel-based toolkit to measure the level of DQ and identify associated DQ issues is developed. By conducting a series of online meetings with two engineering experts from industry, the proposed framework and toolkit have been successfully validated with the help of a case study. The industry experts unanimously concluded that the framework and toolkit could increase the awareness and knowledge of DQA processes, providing good level of understanding and applicability. However, their implementation requires further investigation and in-depth expected benefits identification.

This study contributes to the DQ research by demonstrating a structured DQA framework using an ontology-based approach for capturing the knowledge obtained from the assessment. The DQ ontology is shareable and reusable, unifying the representation of DQA domain knowledge. Future work would include further application of the framework and improvements according to feedback to enhance its applicability and flexibility. Additionally, further work could enable automated knowledge extraction from the ontology, appraise the costs and benefits of the framework and explore machine-learning techniques to improve DQ.

Acknowledgements

This research is supported by the Centre for Digital Engineering and Manufacturing at Cranfield University. The authors would like to acknowledge Babcock International for supporting this work.

References

[1] Ehrlinger L, Wöß W. A Survey of Data Quality Measurement and Monitoring Tools. Front. Big Data 2022; 5:850611.
[2] Heinrich B, Hristova D, Klier M, Schiller A, Szubartowicz M. Requirements for Data Quality Metrics. J. Data and Information Quality 2018; 9:1-32.
[3] Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal 2015;14:2.
[4] Moore S. How to Create a Business Case for Data Quality Improvement 2018.
[5] Cichy C, Rass S. An Overview of Data Quality Frameworks. IEEE Access 2019;7:24634-24648.
[6] Wang R.Y, Strong D M. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems 1996;12:5–33.
[7] Ehrlinger L, Werth B, Wöß W. Automated continuous data quality measurement with qualIe. Int. J. Advanced Software 2018;11:400–417.
[8] Askham N, Cook D, Doyle M, Fereday H, Gibson M, Landbeck U, Lee R, Maynard C, Palmer G, Schwarzenbach J. The six primary dimensions for data quality assessment. DAMA UK group 2013:432-5.
[9] Woodall P, Oberhofer M, Borek A. A classification of data quality assessment and improvement methods. Int. J. Information Quality 16. 2014;3:298-321.
[10] Ardagna D, Cappiello C, Samá W, Vitali M. Context-aware data quality assessment for big data. Future Generation Computer Systems 2018;89:548-562.
[11] Sebastian-Coleman L. Measuring data quality for ongoing improvement: a data quality assessment framework. Newnes; 2012.
[12] Pipino L, Lee Y, Wang R. Data quality assessment. Communications of the ACM. 2002;45:211-8.
[13] Bergdahl M, Ehling M, Elvers E, Földesi E, Körner T, Kron A, Nimmergut A. Handbook on data quality assessment methods and tools. Wiesbaden. 2007.
[14] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM CSUR. 2009;41:1-52.
[15] Camera F, Erkoyuncu J A, Wilding S. Service Data Quality Management Framework to Enable Through-life Engineering Services. Procedia Manufacturing. 2020;49:206-210.
[16] Myers D. About the Dimensions of Data Quality. 2017.
[17] Debattista J, Lange C, Auer S. daQ: an ontology for dataset quality information. Central Europe Workshop Proceedings 2014;1184. CEUR-WS.
[18] Johnson S G, Speedie S, Simon G, Kumar V, Westra B L. A data quality ontology for the secondary use of EHR data. AMIA 2015 Annual Symposium Proceedings. 2015:1937-46.
[19] Nayak A, Božić B, Longo L. Linked Data Quality Assessment: A Survey. In: Xu C, Xia Y, Zhang Y, Zhang L J (eds) Web Services – ICWS 2021. ICWS 2021. Lecture Notes in Computer Science 2022;12994. Springer, Cham.