Baltic J. Modern Computing, Vol. 8 (2020), No. 3, 391-432
https://doi.org/10.22364/bjmc.2020.8.3.02

Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment

Anastasija NIKIFOROVA

Faculty of Computing, University of Latvia, 19 Raina Blvd., Riga, LV-1586, Latvia

Anastasija.Nikiforova@lu.lv

ORCID: 0000-0002-0532-3488

Abstract. Data quality issue has emerged since the end of the 60's, however, more than 50 years later, it remains unresolved and is still current, mainly due the popularity of data and open data. The paper proposes a data object-driven approach to data quality evaluation. This user-oriented solution is based on 3 main components: data object, data quality specification and the process of data quality measuring. These components are defined by 3 graphical DSLs, that are easy enough even for non-IT experts. The approach ensures data quality analysis depending on the use-case. Developed approach allows analysing quality of "third-party" data. The proposed solution is applied to open data sets. The result of approbation of the proposed approach demonstrated that open data have numerous data quality issues. There are also underlined common data quality problems detected not only in Latvian open data but also in open data of 3 European countries.

Keywords: data object, data quality, domain-specific modelling language, open data.

1. Introduction and motivation

ISO 9000 defines "**quality**" as a degree to which the consumer's needs are satisfied, by representing all the characteristics of the product or service requested by the customer. The concept of "**data quality**" derived from the concept of "quality" is usually defined as the suitability of the data for use case, emphasising its relative and dynamic nature, the context of which is determined by the data use and the requirements that depends on it and may change over time, that is determined by data gradual accumulation in the databases, and changing data quality requirements.

Data quality issue has emerged since the end of the 60's, when some aspects of it were first studied by statistical researchers. Computer scientists began actively studying the data quality problem in the early 90's (Scannapieco et al., 2002a). However, despite the popularity of the data and the continued increase in their volume, almost 30 years later, the data quality problem remains unresolved and is still current, mainly due the popularity of [open] data. An analysis of more than 70 existing solutions reveals that most existing solutions are based on the definition, grouping of data quality dimensions

and their application to data sets, that are often identified by researchers as problematic even for data quality professionals. Therefore, there is reason to argue that existing approaches are not suitable for users without in-depth knowledge on IT and data quality, so that the involvement of data quality specialists becomes necessary at all stages of the data quality analysis. Nowadays, this is not acceptable, because every day users come into contact with data – they are everywhere, so that the ability to verify their quality must be for each user, regardless of his/ her knowledge in the field of IT and data quality.

The paper proposes a data object-driven approach to defining and evaluating data quality. The aim of the study is to define the developed approach that would allow to define the data object to be analysed, as well as quality requirements for people who may not have in-depth knowledge in the fields of IT and data quality, and to apply it to data sets by demonstrating it in action.

The proposed data quality model consists of three components: (1) **a data object** whose quality is assessed (primary and secondary for contextual data quality analysis), (2) **data quality specification** – quality requirements defined for a previously defined data object, which depends on a specific data usage, and (3) a data quality verification process which results in a determination of the quality of the given data object by analysing the identified data quality issues. The proposed approach differs significantly from existing approaches - it does not use the concept of "**data quality dimension**", allowing users to define their own specific quality requirements for their specific data objects, depending on the use-case. Instead of the concept of "data quality dimensions", a broader concept of "**data quality requirement**" is used, where "data quality dimension" may be considered to be a subset of "data quality requirement" concept. The data object and quality requirements for a particular data object are defined by the user, whereby users are given the opportunity to verify the quality of the data of a specific dataset for their purposes. Each component is defined by using graphical flowchart-like charts that make it easier to analyse data quality, and by ensuring the interaction between users through charts that can be created and edited quickly and easily. This is achieved by developing a graphic domain-specific language (DSL) for each component. A quality model can be defined in two ways, informally using a natural language, or formally, by replacing non-formal texts with executable, such as SQL queries. The definition of data object and data quality requirements does not require users to have prior knowledge of IT or data quality; this process is intuitive, whereby, unlike a larger part of existing data quality solutions, the proposed approach is intended for a wide range of users. The involvement of IT specialists is only becoming necessary at the final stage - transforming informal requirements into executable.

The proposed solution makes it possible to carry out a quality analysis of "third party" data, i.e. to analyse data, information on which storage and processing mechanisms or procedures are not known. The solution is applied to open data verifying the effectiveness of the approach as well as the quality of the open data, placing more emphasis on open data in Latvia. An analysis of the quality of open data by itself is a challenge, since despite the increased popularity of open data, the issue of open data quality is studied relatively rarely, as is also the case with the number of studies presenting *Google Scholar* on the relevant topic - the studies on open data quality are carried out unjustifiably rarely. As the volume of open data increases, solutions are needed that would be suitable for users without more in-depth knowledge of data quality and IT, as open data becomes a daily phenomenon and quality analysis is becoming an

integral part of everyday activity. As a result of the analysis, a number of data quality problems were identified in open data, which, in view of their nature, were divided in separate groups in order to draw attention to the common problems from which data users should be aware, and to take into account data providers, highlighting the most popular ones.

The objectives of this paper:

1) to explore the concepts of "*data quality*" and "*open data quality*", challenges related to these concepts, their relevance and popularity in scientific articles;
2) to impose requirements for a new data quality assessment approach and to propose a new approach for data quality analysis and evaluation, that meets the identified requirements;
3) to assess the proposed data quality assessment approach by applying it to a number of open datasets in order to identify data quality problems in them.

The paper is structured as follows:

- Section 2 provides definitions of basic concepts, a description of the relevance of data quality and a justification for the problem. It is also explores the popularity of data and open data-quality studies based on *Google Scholar* data, identifying topics that are covered in the scientific studies more rarely, and which research could bring added value to society;
- Section 3 briefly discusses existing solutions and studies on data quality issues, defining their common negative features, which would be worth taking into account developing a new approach;
- Section 4 presents a proposed new approach to the definition and assessment of data quality, which addresses the disadvantages of existing solutions. The selection of its components is justified, highlighting its advantages compared to existing solutions. A data quality analysis is described for both a single data object and multiple data objects corresponding to contextual data quality analysis. Since the data quality model can be defined both informally and formally, which is in line with the basic design of a model driven architecture (MDA), the proposed solution is examined from the MDA perspective by analysing their relevance to MDA's basic ideas, highlighting both similarities and differences;
- Section 5 summarises the results of the application of the proposed approach to the real data sets, identifying common data quality problems;
- Section 6 sums up the conclusions of this research.

2. The concept of data quality

The concept of **data quality** has several definitions, but the most often this is defined as fitness for use (e.g., (Tayi et al., 1998), (Olson, 2003), etc.). Some researchers add to this the requirement that data must have no data quality problems and must have the necessary or "desirable" properties (Scannapieco et al., 2002a), (Redman, 2001), (Wang et al., 1996). These requirements may vary from solution to solution. For example, Lee and co-authors believe that qualitative data is characterized by features such as *completeness*, *consistency*, *believability*, *timeliness*, *accessibility*, adding that the data must be *available in an appropriate amount* and *free of error* (Lee at el., 2009). The list

of properties that Juran proposes to use evaluating the quality of data overlaps (Lee et al., 2009). According to (Juran, 1995), the data must be *available*, *accurate*, *timely*, *complete*, *consistent with other sources*, *relevant*, *comprehensive*, *proper level of detail*, *easy to read*, *easy to interpret*, etc. However, the lists of properties or data quality dimensions that characterize data quality are very diverse.

In general, data quality is the suitability of a given dataset and its properties for a particular usage or use case, which depends on the data consumer using them, for example, in analytics, making business decisions, planning etc. (Nikiforova, 2018b). It also means that the same data may be suitable for one application or user but unusable due to low quality for another (Tayi et al., 1998). This means that it may be necessary to define different data quality requirements for the same data, depending on the use case.

It also means that it is not possible to achieve a level of data quality at which the data would satisfy all possible use cases, more precisely the **absolute data quality**, however, this is the objective to be pursued. This principle is common for many quality-related topics, such as the QMS (Quality Management System), as many methodologies, such as LEAN and its descendants, for which one of the main objectives is to improve the efficiency of the business process, which posse this principle in the core of philosophy (Nikiforova and Bicevska, 2018). It should also be noted, that, despite the main challenge of project failure is usually associated with the incorrect selection or non-use of project management methodologies, significant part (around 40%) of business initiatives fail due to insufficient quality of data (Friedman and Smith, 2011).

In the 21st century, a new concept derived from the concept of "data" emerged – "**open data**", bringing new challenges arising from their nature. This topic is discussed in the next subsection.

2.1. Open data

The popularity of the data quality problem is growing more rapidly, largely due to the popularity of open data. Nowadays, countries are enabling users to obtain data from open data portals where different types of data are published by various data providers. These data can be used by data users for their own purposes, ranging from simple data analysis for their own purposes, to the trend of developing applications based on open data today.

Open data is the preferred way of making available data re-usable. For data to be recognized as open data, they must be: (1) *complete*; (2) *primary*; (3) *timely*; (4) *available*; (5) *machine readable*; (6) *non-discriminatory*; (7) *non-proprietary*; (8) *license-free* ((Bauer et al., 2011), (WEB, e)). However, none of the open data principles are related to data quality in its broad meaning, therefore there is reason to assume that open data (even those that fully comply with all the above principles) tend to have data quality problems (Nikiforova, 2018b).

The same trend can be observed in the case of open government data (OGD) evaluation, as according to (Klein et al., 2018), the quality aspect takes only the 4th position (out of 4) by popularity, following the policy, benefit and risk, despite the fact that quality can affect every aspect. However, according to https://www.europeandataportal.eu/en research, data quality is the most problematic aspect of open data portals.

It should be also stated, that since open data portals publish data from different data providers, the quality of data sets within a single open data portal tends to vary, which is

also consistent with ((Kuk et al., 2011), (Petychakis et al. , 2014)). This tends to be related to the fact that, when publishing datasets, data portals rarely check their quality, which is usually related to the complexity of the data quality review process. As a result (Kuk et al., 2011) and (Yi, 2019) conclude that nowadays the quality of OGD data is not high enough and quality problems often occurs, starting with irregular data updates and incorrect format selection, which mostly relates to the quality of data sets quality problems, continuing with data incompleteness, name and identifier contradictions, low level of granularity etc..

2.2. Data quality problem and its popularity

The data quality issues were firstly researched by statisticians in the late 60's. At the beginning, mainly mathematical theory for detecting and eliminating duplicates in statistical datasets was proposed. In the early 80's, studies on data quality management were launched, focusing mainly on identifying and tackling data quality problems in the management solutions for production systems. In the early 90's, computer scientists have also studied this problem, focusing on defining the concept of data quality, measuring and improving the quality of data stored in databases, data warehouses and legacy systems (Scannapieco et al., 2002b), as well as linking "data quality" concept to "data quality dimensions", proposing different dimension groupings (Cai et al., 2015). However, despite the popularity and continued growth of the data (Hashem et al., 2015), (Kitchin, 2014), (Cai et al, 2015), almost 30 years later, the quality problem remains unresolved (Cai et al., 2015).

Nowadays, a variety of studies are carried out each year, including estimates and surveys aimed at identifying the effects of data quality, including losses caused by low quality data. The results raise awareness of the topics and call for solutions to improve the current situation. Results of some surveys and analyses:

- in 2018, low-quality data was considered "the leading cause of failure for advanced data and technology initiatives, to the tune of $9.7 million to American businesses each year" (WEB, c);
- annual Gartner Group research (Moore, 2017) demonstrates that companies lose $15 million annually due to data quality problems. Moreover, this trend has been constant in recent years (e.g. (Friedman et al., 2013), (Moore, 2017));
- IBM's research (WEB, d) found that business decisions taken on the basis of low-quality data cost the US economy $3.1 trillion per year;
- the US postal service provider USPS has a loss of US $3.4 billion per year due to incorrect address data (WEB, b).
- according to Gartner studies, about 40% of data in companies is of poor quality, while data quality is closely linked to process quality and, as a result, to business success (Friedman et al., 2011).

In some cases, data quality problems bring financial loses, however there might be cases, when they bring even more global outcomes. The two most impressive examples are the explosion of the space shuttle Challenger and the shooting down of an Iranian Airbus by the USS Vincennes where, according to (Fisher et al., 2001), data quality problem was one of the crucial. According to them, data consistency, completeness and accuracy were the most critical aspects.

In accordance with ((Jetzek et al., 2017), (Chen et al., 2016), (Colpaert et al., 2013)), the quality of the [open] data affects the quality of knowledge, its reliability and the significance that can be gained from processing the data.

According to Loshin (2001), low quality data reduces the efficiency of work, so when working with data, one must be sure that they are correct before being added or processed - they must be correct at every stage of data accumulation. If the data has an error, it must be corrected, or the record must be deleted before it can be used. The more errors that have accumulated, the more resources are needed to fix them. Low-quality data also affects business decision-making, while high-quality data improves the efficiency of data warehousing, as data retrieval, cleaning, and downloading typically take up to 80% of the time. It is also consistent with (Gabernet et al., 2017), according to which the widely used '80-20 rule' is also applicable to data quality, whereby 80% of the time of a data researcher 20% leaving it for use, including analysis.

All these numbers demonstrate that the data quality problem is current, since the quality indicators are unsatisfactory and, unfortunately, constant. The existence of a quality problem is also confirmed by a number of studies (Acosta et al., 2013), (Färber et al., 2016), (Ferney et al., 2017), (Guha-Sapir et al., 2002), (Kerr et al., 2007a, 2007b), (Kontokostas et al., 2014), (Kuk et al., 2011), ((Nikiforova, 2018a, 2019), (Vetrò et al., 2016), (Yi, 2019) etc.).

The relevance of the [open] data quality problem and its popularity is also reflected in the number of studies. According to *Google Scholar*, the number of studies on open data quality published between 2003 and 2014 is 4.6 times fewer than in 2018. The results of the research show that a sharp increase in the popularity of open data quality has been seen since 2017, when the number of open datasets and the number of open data portals has started to increase (Figure 1). However, relating the number of studies related to open data quality to the total number of studies related to open data, it appears that the data quality issue has been studied unjustifiably rarely (Figure 2), since the number of studies on open data in 2018 exceeds the number of open data quality studies at 147 times (in 2019 – 179 times), i.e. the proportion of open data quality studies against the total number of studies related to open data shall not exceed 0.5% (the data were collected in the 3rd quartile of 2019).

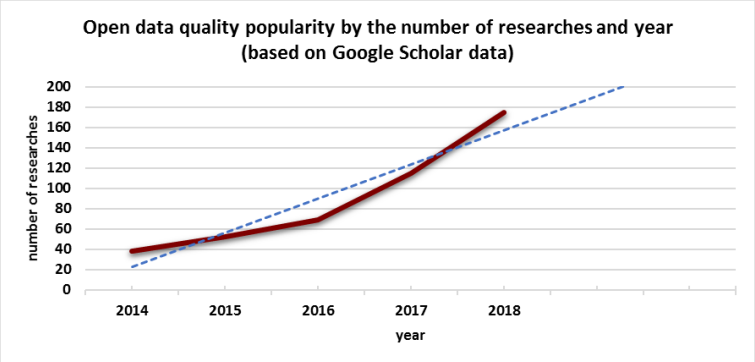

Fig. 1. Popularity of open data quality studies (2014-2018)

In addition, the number of data quality studies exceeds the number of open data quality studies at nearly 196 times (i.e. the ratio of open data quality studies against the total

number of data quality studies, is ~ 0.2%). This shows the need to carry out studies related to the quality of the open data, as the results of data quality studies show the existence of a problem, but the distribution of the studies carried out shows that the existing solutions are mostly designed to assess the quality of the so-called "closed" data and are not suitable for open data or users without advanced IT and data quality knowledge.

At the same time, the topic of the quality of [closed] data is losing popularity (Figure 2), while the popularity of open data and open data quality topics is steadily increasing (Figure 1).

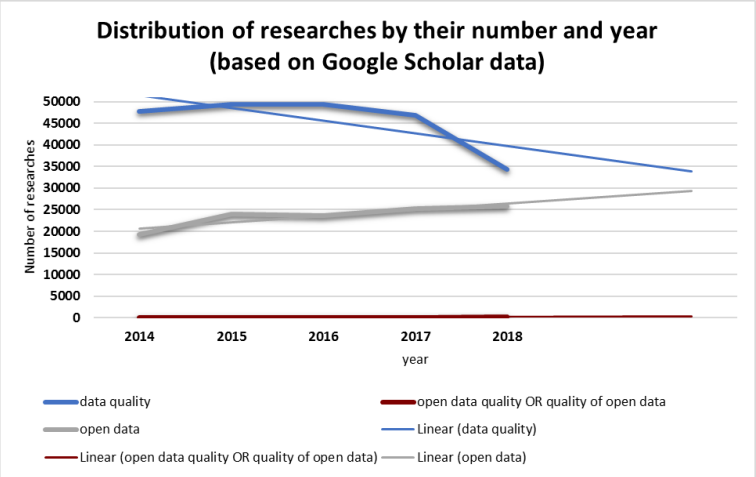

Fig. 2. Popularity of data quality studies (2014-2018)

As the volume of open data increases, solutions suitable for users without in-depth knowledge on data quality and IT become necessary, as open data becomes a daily phenomenon and their quality analysis becomes an integral part of everyday activity.

3. Existing solutions

As the topic of existing studies was addressed in (Nikiforova, 2019) and (Nikiforova et al., 2020), this discussion will not be repeated, emphasizing only the main categories of existing studies:

- general studies on data and information quality, mostly focusing on definition of data quality dimensions and their groupings ((Wang et al., 1996), (Van den Berghe et al., 2017), (Ferney et al., 2017), (Redman et al., 2001) etc.);
- quality assessment of open data portals and/ or Open Government Data ((Vetro et al., 2016), (Kučera et al., 2013), (Nikiforova, 2020a, 2020b), (Neumaier et al., 2016), (Sáez Martín et al., 2016), (Sasse et al., 2017) etc.);
- quality assessment of Linked Data ((Acosta et al., 2013), (Paulheim et al., 2014) etc.).

Some studies conduct industry-specific data and information quality analysis, most often using sector-specific methods that are tailored to these datasets:

- cancer registry ((Bray et al., 2009), (Sigurdardottir et al., 2012), (Larsen et al., 2009), (Parkin et al., 2009), (Tomic et al., 2015) etc.);
- healthcare ((Dahbi et al., 2018), (Weiskopf et al., 2013), (Van den Berghe et al., 2017), (Schmidt et al., 2015), (Kerr, 2007a, 2007b) etc.);
- chemical hazard and risk assessments ((Bevan, 2012) etc.).

It should be noted that some studies may belong to several groups at the same time as these groups are interconnected and several groups could be subdivided, for example, individual studies also develop guidelines for data quality (e.g. (Aarshi et al., 2018), (Kucher et al., 2013), (Vetro et al., 2016), (Sasse et al., 2017), (Perez-Castillo et al., 2018a, 2018b) etc.), but these are usually derived from the results of the study.

Definition of data quality dimensions and methods for their quantitative evaluation is one of the most important steps that were taken so far in the field of data quality (Nikiforova, 2018a). Current data quality analysis solutions are largely focused on informal definition of data quality and measurement of acquired values, but mechanisms for determining data quality characteristics in formalized languages are unknown (or popular enough). Similarly, there are no well-known solutions that allow users to simply analyze the quality of specific datasets by defining specific data quality requirements for individual parameters of interest (Nikiforova, 2018a).

Summarizing the previous chapter, it should be emphasized that “data quality” is a complex concept that depends on the particular application of the data. The use of data quality dimensions in data quality analysis can appear complicated, since despite the age of data quality analysis studies that relates the concept of data quality to the concept of data quality dimension, it is still unclear how and which specific data quality dimension to associate with a particular use-case. This lack of existing solutions is also pointed out by several data researchers, including Batini, author of an in-depth study of data quality issues and existing methodologies (Batini et al., 2009, 2016). However, this is confirmed not only by the numerous data scientists but also by the survey of computer science students (55 respondents in total) who were asked to define the concept of “data quality”. As a result of the survey, several identical definitions were not found since even with a similar definition, they were supplied with a list of different characteristics that, in the respondents' opinion, describe the quality of the data. Naming three most important dimensions of data quality, names of only 9 dimensions were mentioned more than once, namely, "*accuracy*", "*relevance*", "*integrity*", "*duplication*", "*accuracy*", "*availability*", "*unambiguity*", "*trustworthiness*", "*completeness*". In addition, 10% of respondents couldn’t name any of them due the fact that the concept of “data quality dimension” in the context of data quality isn’t known for them at all, while 12% of respondents provided the list of concepts that can’t be accepted as data quality dimensions, therefore, it can be concluded they haven’t deal with data quality dimensions, too. Among the 78% of respondents, the most frequently mentioned dimension is "*completeness*" – this dimension occurred for 5 times, "*accuracy*" - 4 times, while the other dimensions were mentioned by 2 to 3 respondents. However, it should be noted that despite having the same name for the named dimensions, it cannot be stated that respondents understand them in the same way. It should be noted that, despite the fact that third-year undergraduate students in Computer Science who might be observed as users with in-depth knowledge in IT area, took part in the survey, their knowledge of data quality issue, and in particular the data quality dimensions, is very limited. In

addition, 55 non-IT experts were surveyed to find out whether the concept of "data quality" is known for them, how it could be defined and what the data quality dimensions might be associated with it. The survey showed that (1) 96.4% have not previously heard the term "data quality dimension", (2) only 7.3% of respondents' assumptions about the definition of this term are correct and 17.1% are partly correct, (3) 16.4% of respondents could name at least one existing data quality dimension. This proves once again that the linking of the concept of "data quality" to "data quality dimension" by the end-users without in-depth knowledge of data quality should be considered as very risky task.

Most of existing solutions are not suited for non-IT and non-data quality experts, since they require additional in-depth knowledge not only in IT but also in data quality area, especially if one of ((Caro et al., 2007), (Ferney et al., 2017), (Neumaier, 2015), (Umbrich et al., 2015), (Vetro et al., 2016)) is used. These solutions are suited for users with appropriate knowledge, skills and experience in data quality area or involving such users at all stages of data quality analysis since they (a) use high number of data quality dimensions, complicating this task significantly, (b) require definition of data quality requirements, (c) require linking defined or pre-defined data quality requirements to data quality dimensions, which further are applied to datasets. The involvement of experts at all stages of data quality analysis is inappropriate since this contradicts the main principles of data quality – data conformity to use-case which must be defined only by the end-user who is analyzing the quality of data for his own purposes. Their involvement is acceptable on the few of later stages only, however both, data under analysis and data quality requirements against which data quality will be analyzed, must be defined by end-user, while IT-experts are allowed to support data quality analysis.

The nature of existing studies requires the division of users involved in data quality analysis based on the knowledge required for this task into two groups: (1) IT specialists and (2) data quality specialists (an overlap of both sets is possible). The distinction between IT and data quality professionals is necessary because (1) IT professionals may not have data quality knowledge, i.e. knowledge necessary for data quality analysis, (2) a user may be considered a data quality expert if his / her qualifications and / or experience are sufficient to perform data quality analysis, whereas this knowledge may be available to a user who does not have advanced IT knowledge. The second group can be represented by data analysts working in banks or other fields with a sufficient level of knowledge in data quality analysis, despite the lack of IT training. This means that a user can be considered a data quality specialist if he / she has an in-depth knowledge of data quality concepts and is capable of performing the above tasks related to data quality analysis (Nikiforova, 2018a). According to (Nikiforova, 2018a), if the data quality analysis solution requires in-depth IT knowledge (e.g., (Acosta et al., 2013), (Färber et al., 2018), (Kontokostas et al., 2014), (Redman, 2001), (Zaveri et al., 2016)), the concept of IT specialist applies. A user is considered an IT professional if he or she has education and / or experience in the IT field that covers relevant topics (i.e., specific technologies, approaches, knowledge engineering, etc.).

It should be noted that despite the diversity of existing solutions for data quality analysis, the users or groups of users involved in them, the knowledge or technology needed, etc., a common feature of data quality studies is the identification of the presence of data quality problems in the data to which the proposed solutions are applied. This means that the problem of data quality is still unsolved and (a) further research is needed in this area, proposing new solutions; (b) existing [open] datasets

should be examined, providing information on the weaknesses identified, thereby improving their quality. To achieve this aim, a new approach to data quality analysis was developed, which is discussed in the next section.

4. The proposed data object-driven approach to data quality evaluation

This Section defines requirements for a data object-driven approach to data quality analysis, describes the proposed data quality model, providing an overview of each of its components. It also provides a description of the possibility of contextual data quality analysis, highlighting the advantages of the proposed approach in comparison with existing solutions. A general description of the approach is given, its description in the context of model driven architecture (MDA), justifying the choice of components and their combination in the overall solution. The Section concludes with a list of opportunities and limitations of the proposed approach. This Section is mainly based on (Nikiforova et al., 2020), (Nikiforova and Bicevskis 2019).

4.1. General overview of the approach

Taking into account the relative and dynamic nature of the data quality, according to which data quality requirements are determined by data use-case, specific data quality checks may be required for each specific application. This also corresponds with (Batini et al., 2009), in which authors emphasize that the data quality solution is based on three aspects: (1) data and process analysis, (2) data quality requirements analysis, and (3) data quality analysis. In addition, since data is usually collected gradually, the following basic requirements for a data quality management system are:

- in accordance with the dependence of the concept of data quality on the use case, data quality requirements should be formulated in platform independent concepts, i.e. not including checks in the IS program code;
- data quality requirements should be formulated at several levels, i.e. for a single data object, a data object in the context of its attributes, a data object in the context of a database, a data object in the context of many IS;
- the language used to define components of data quality model should be simple enough, ensuring possibility to define data objects and data quality requirements even for industry professionals with minimal involvement of IT specialists. In accordance with (Zhao et al., 2003) this can be achieved by using graphical DSL, the syntax and semantics of which are easily adapted to each new IS;
- data quality must be verified at several stages of data processing, each time using its own description of individual data quality requirements. It is advisable not to try to include all data quality requirements into one comprehensive requirement specification that would only test data quality at the final stage of data collection, i.e. when the data is already stored in the database.

From the Batini and co-authors perspective (Batini et al., 2009), the proposed approach intends maximizing user engagement rather than questioning it, where the end

user is only one of the data quality analysis stakeholders, allowing the user to define every step of data quality analysis.

From TDQM (Total Data Quality Management) point of view, data quality lifecycle consists of 4 interrelated phases - **data quality definition**, **data quality measurement**, **data quality analysis** and **data quality improvement**. According to this, the proposed approach can be briefly described as follows:

- phase 1 – **definition of the data quality**, which includes (1) selecting the data object whose quality is under analysis, (2) defining the data quality requirements for the data object class. It is expressed by a set of conditions whose fulfilment is checked. Data quality requirements are captured by graphical graphs, where the vertices of the graph represent the operations and checks to be performed, while the arcs represent the order in which they are executed. Data quality requirements can be formulated at different levels of abstraction, from informal text (for example, in natural language) to precise, executable program artefacts, replacing informal texts with executable code or SQL queries.
  With regard to reading data from data sources, it should be noted that the use of the term "data object" implies that only data necessary for a specific analysis is selected, thus reducing the amount of data to be processed, saving time and other resources;
- phase 2 - **data quality measurement phase** that intends that (1) the data to be analysed is selected from data sources (screen fields, databases, files, data warehouses, etc.) and (2) data quality measurements are performed by verifying the fulfilment of previously defined requirements for each data object. One quality measurement process can include reading multiple data objects and testing requirements with multiple quality specifications. The quality assessment process results a protocol containing non-conformities with the quality specification identified during the inspection process;
- phase 3 - the data quality analysis phase, that intends that the analysis of the data quality test results received during the measurement phase is performed. Its purpose is to identify data quality problems and to identify the root causes of these problems;
- phase 4 - data quality improvement phase that involves the selection and implementation of a quality enhancement mechanism. This can be done both with customized software modules and with the tools provided by MS DQS (*Microsoft SQL Server Data Quality Services*), a data quality analysis and enhancement tool whose pros and cons and potential applications have been addressed in (Nikiforova, 2018b), hence the proposed solution does not address this phase.

Since, according to TDQM, data quality can be ensured by systematically repeating the phases of the data quality cycle, that is necessary since data is constantly changing, and new or modified data may lead to new data quality problems or changes in data quality requirements. Therefore, the proposed solution ensures possibility to change data quality criteria, defining new data quality requirements for each new iteration cycle, thus ensuring and maintaining high data quality.

The data quality management system architecture is illustrated in Figure 3. The main components of the proposed approach, i.e. proposed data quality model are:

1) **data object** that defines the data whose quality will be analysed;

2) **data quality requirements** - conditions that must be fulfilled for data to be considered as of high quality;
3) **data quality evaluation process** - all activities that must be performed to evaluate the quality of the data object.

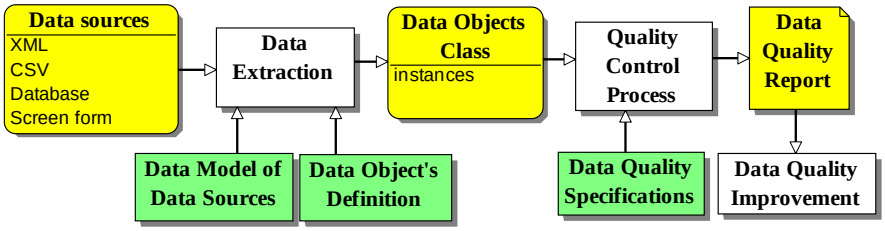

Fig. 3. Architecture of data quality system [(Bicevskis et al., 2019a) modified by the author]

All three components of the data quality model are described by language metamodels. In addition to syntax, they also instruct the graphical representation of the model, while the semantics of the model are described by the execution rules of graphical diagrams. This approach corresponds to (Zhao et al., 2003). In this way, data quality models become executable and practically usable.

In accordance with widely-spread data quality division into **syntactic** and **semantic accuracy**, and many problems arising from the use of the data quality dimension concept, the proposed data object-driven approach abandons the data quality dimension concept by replacing it with the more universal and broad concept of "data quality requirement", where "data quality dimension" related to the quality of the data is a subset of it. In other words, the concept of data quality is not related to the concept of data quality dimension. As follows from Chapter 2, understanding the concept of a data quality dimension, as well as defining specific dimensions with approach developers and exploring with users of a particular approach, is too resource consuming. This has been repeatedly acknowledged by even the brightest data quality researchers (Batini, Scannapieco, Shanks, etc.). In addition, as mentioned in Section 2, some studies use too many dimensions of data quality, while others limit it to two dimensions. The proposed approach does not set any limits, giving users the ability to define data quality requirements that depend on the use case. In scope of the given research, the use of the term "data quality requirement" instead of "data quality dimensions" has several and notable benefits since, firstly, it enables the involvement of users, who may not have advanced knowledge of IT and data quality, in data quality analysis, secondly, it facilitates multi-user collaboration; and third, does not limit the nature of the quality requirements to be set (compared to data quality dimensions, where each specific dimension may have a limited list of possible data quality checks depending on its implementation). In addition, it saves developers and users time simplifying the process of designing and using data quality solutions without requiring a number of resource intensive activities related to the data quality dimensions, i.e. from the developers' point of view, defining the dimensions of data quality with all its consequences, including grouping them, selecting attributes, metrics, selecting and providing measurement mechanisms, etc., from the user's perspective, learning all the dimensions and components of a solution that are usually useful in only one particular solution.

However, despite abandoning the concept of the data quality dimension by replacing it with the more universal one, the proposed approach follows the generally accepted definitions of the concept of data quality.

As for data quality assurance areas, i.e. for a single data object, for a data object in the context of its attributes, and for a data object in the context of a database that relates to different IS components, the proposed strategy allows to propose a unified solution. This is achieved by offering a DSL platform that provides a wide range of language definitions for describing data quality requirements, allowing modular definition of quality requirements, and verifying compliance with requirements at various stages of data processing, i.e. without including all requirements in one comprehensive requirement specification, where data quality is checked only at the final stage of data collection. This enables the quality of the data to be verified at many stages of data processing, each time using its own description of individual data quality requirements.

Another idea of the proposed data object-driven approach is the involvement of data users in data quality analysis. It is vital not only for the users themselves, who must be able to analyze the quality of the data for their own purposes, but also for the data providers, as the increasing amount of open data published (number of datasets) results a high number of different possible use-cases, that data publishers are unable to put forward and verify, however, user engagement in data quality analysis and their feedbacks would increase the quality of potential data significantly, increasing the probability that as many data quality defects as possible will be identified and ultimately eliminated ((Ruijer et al., 2019), (Attard et al., 2015), (Tinholt, 2013)). The proposed solution allows evaluation of data quality according to objective and subjective requirements, where **objective requirements** are user-independent requirements that allow to evaluate data compliance with pre-defined requirements, integrity laws, or external sources (Price et al., 2005). By contrast, **subjective requirements** mean data-dependent requirements that tend to change depending on the task that requires the data at a given time. They are highly dependent on the perception of the data user of the concept of "data quality".

The proposed data quality model can be formulated and used in at least two ways / levels: (a) informally (similar to PIM), where the required checks are described in natural language - chart symbols contain textual descriptions of the actions; (b) in an executable manner (similar to PSM) that can be achieved by transforming an informal model, replacing informal texts / descriptions with program code, SQL queries, or other executable objects. PIM can be seen as an informal description of an IS that is created with industry professionals, i.e. users who may not have advanced IT knowledge. In such a way the model is gradually detailed.

This means that the proposed data quality model divided into two models can be described from the point of view of the Model-Driven Architecture (MDA). Perhaps it is not MDA in its traditional meaning, however, the principles are the same. According to (Kleppe, 2003), MDA by itself "… *is based on widely used industry standards for visualizing, storing and exchanging [software designs and] models*". This is the core idea of the presented approach. Following Object Management Group (OMG) (Soley et al., 2000), in the proposed solution "*models become assets instead of expenses*", as well as one of the main objectives of using the charts is *"modelling technology to pull the whole picture together*". All concepts of the presented data object-driven approach are defined and described using graphical Domain Specific Languages (DSLs). DSL syntax and semantics are developed in such a way that they are (a) easily applicable to a new IS,

(b) simple enough to let non-IT experts define data object and quality specification without IT-experts involving. Graphical models for data quality analysis were chosen for several reasons. Firstly, models are usually used as a communication tool (Mellor et al., 2004), improving the readability of information since graphical representation in models is perceived better by readers than textual representation. Visual information is also easier and faster to read and to modify. The use of models reduces the risk of misunderstandings between users. According to (Mellor et al., 2004), models are "*cheaper to build than the real thing*". Mellor also emphasizes that the effectiveness of models depends on two aspects: abstraction and classification. By **abstraction** Mellor understands "*ignoring information that is not of interest in a particular context*". In the presented approach, it is achieved by using data object exclusively with the parameters representing real objects that are of interest for specific users in specific use-cases. Parameters that are not of interest for specific use-cases are ignored, hence they are not included in the particular data objects. By **classification** Mellor means "*grouping important information based on common properties*". This principle is partially followed when grouping quality conditions for each parameter involved in data quality analysis. In (Kleppe et al. 2003), the authors propose to create machine-readable models instead of the paper-based to reduce time- and effort- consuming activities. They offer to store machine-readable models in standardized repositories. In the presented approach, a graphical DSLs editor DIMOD is used to store created diagrams in repository. It should be noted that a similar approach, i.e. the use of graphical models for data quality analysis tasks, is also used by one of the leading data quality researchers, Scannapieco (Scannapieco et al., 2002b), where authors emphasize the lack of modelling languages intended for data quality [improvement]. The authors point out that the modelling language suitable for data quality improvement tasks must be formal enough to ensure a unique and unambiguous interpretation of language structures. In the research, the authors point to the need for a language that would be simple enough to be used by users without in-depth knowledge of IT, emphasizing that interaction with the end-user is a primary challenge for data quality analysis. Unlike authors who have preferred Unified Modelling Language (UML), which also meets the vision of OMG, the proposed solution prefers flowchart-like graphical DSL.

While UML charts are the most commonly used modelling technique in MDA (Kleppe et al., 2003), which is also used by Scannapieco (Scannapieco et al., 2002b), taking into account that UML diagrams often require specific knowledge and previous experience, UML is one of the most appropriate choices for engineers, as it allows exchanging and documenting their ideas (comply with Kleppe et al. (2003)), however, UML is not suitable for non-IT and non-DQ experts, therefore it can't be used for the proposed solution. In addition to traditional UML, i.e. without extensions, it is considered too superficial and general (Scannapieco et al., 2002b). Because of UML's shortcomings, in (Haubold et al., 2010) UML is used in combination with DSL, the combination of which, in the authors' view, gives better results in addressing each technology's shortcomings (as for DSL - availability was considered as the most important limitation). However, combining UML with DSL greatly facilitates the creation of metamodels (Haubold et al., 2010). At the same time, flowcharts are a simple and intuitive way to express ideas even for non-IT and non-DQ experts, and they are often included in educational programs for secondary schools (at least in Latvia). As a result, author supposes flowcharts are easy to create, read and modify for the majority of users because they have all necessary components for data quality analysis. This makes

it possible to assume that such charts can be easily designed, edited by non-IT-specialists, and will facilitate the communication between the individuals involved in the process of analysis of the quality of data. For these reasons, flowchart-like charts were chosen as the most appropriate option for the proposed solution.

In the light of all the above, a data quality model consisting of graphical models was established where each chart describes a specific stage of data quality evaluation. All checks for one business process are combined into packages, while all packages together form a data quality model. Each chart consists of vertexes and arrows arcs: (a) the vertexes identified by mnemonic graphic symbols, represent the elemental data quality management actions, (b) the arcs connect the vertexes, indicating the sequence of actions to be performed (Nikiforova, 2018a). Other steps can also be included in the charts, such as preparing error reports that are designed to record data quality problems, i.e. creating a protocol that records data that do not meet data quality requirements. The resulting execution protocols are then used to correct the data. Using charts allows users to define a data object and corresponding data which quality will be analyzed, data quality requirements that should be met by data to conclude that they are relevant to a specific task, regardless of their level of knowledge. Describing the requirements in this way excludes the need to describe the requirements in textual form, which may be interpreted differently, thereby also facilitating the realization of the third phase of data quality analysis, i.e. the data quality process, as the possibility of a lack of understanding between the end-user and the data analyst has been excluded or at least significantly reduced.

As programming languages and platforms may have significant differences in their semantics, PIM transformation into PSM takes place manually. Despite there are many options for automated and semi-automated transformation of PIM into PSM, it is almost impossible to ensure the correct translation of PIM defined by users into the PSM. Besides, as it was previously mentioned, the presented solution does not follow MDA in its traditional understanding. One of the main reasons to choose manual transformation is the fact, that the manual transformation of models task isn't effort- and time-consuming in this case – it is relatively simple task, especially for users with basic programming skills which will be required at the later stages of quality analysis only (corresponds with (Lano, 2005), (Miller et al., 2003), (Ostadzadeh et al., 2008), (Pauker et al., 2016), (Carrol et al., 2006), (Chungoora et al., 2013) etc.).

Next subsections are dedicated to components of the proposed model and possibility of contextual data quality analysis.

4.2. Data object

A data object is one of the basic concepts of the proposed approach. A set of parameters that describe a particular real-world object is considered to be a data object. For instance, (1) university and its characteristics - name, registration number, date of establishment, address of the website, contact phone, list of faculties and their names, address, etc. - can be considered as a data object, as well as (2) country and country name, capital, official language, legislature, area, currency, ISO code, list of border states, etc.. One of the most common examples of a data object today is the Wikipedia's info boxes and their content - data about each unit user is searching for. Similarly, a document with completed field values, such as a questionnaire, invoice, etc., can serve as an example of a representation of a data object. They are all joined by the fact that the values of the parameters of the

data object are displayed without coded values, which is typical of storing data in databases. It should be noted that the nature of the data object allows to define it as a result of the process, such as the list of departed routes obtained from the navigator.

The specification for the data object and data quality is based on a use-case, i.e. the purpose for which a particular data object is used – user needs, desires, etc. This means that quality analysis only needs fields that describe real objects that are important to the user and will vary in different cases. As a result, the data object representing one real object may vary depending on the use-case, both in terms of the number of parameters that describe it and in terms of structure (Nikiforova, 2018a).

As part of the study, a data quality analysis of more than 30 open datasets was performed, the results of which are also available in relevant scientific articles (Nikiforova, 2018a, 2018b, 2019a), (Nikiforova et al., 2019). In this paper attention is mainly paid to analysis of (a) one specific domain – quality analysis of Latvia's open health[care] data, (b) the Company Registers of four European countries (Latvia, Norway, UK, Estonia). When analysing data of company register, it is obvious that the data object is "*Company*". In order to give an insight into the given approach, all phases of data quality analysis are examined on a specific example, illustrating the appropriate stage of data quality analysis for the "Companies House UK" dataset. This dataset got the preference since it allows to describe each concept related to the data object. Every company is described using 55 parameters.

Due to the nature of the concept of a data object, the number of parameters depends on the use-case, which corresponds to the principle of abstraction (Mellor et al., 2004). Different use-cases can be defined for a single data object. For example, the results of the survey carried out show that at least 19 different usage examples can be defined for the Company Register. Two very simple and intuitive use-cases were chosen:

1) identify company by its name, registration number and incorporation date;
2) contact company via mail post using its address and postal code (Nikiforova, 2018a).

Defining the same use-cases for all analysed company registers allows to compare the quality of different company registers. Therefore only 5 attributes for the data object "*Company_UK*" are necessary to cover the use-cases: "*CompanyNumber*" – company registration number, "*CompanyName*" – company *name*, "*IncorporationDate*" – company incorporation date, "*RegAddress_ AddressLine1*" – company address, "*RegAddress_PostCode*" – company postal code. Other 50 parameters are out of the scope for this use caseand can be ignored.

In the case of a PIM model, an informal description of the values to be stored (in the natural language) is defined for each parameter. The description of company is informal as no rules for attribute values' syntax are given. The description is non-formal because no syntax conditions are defined for its attributes.

The description of the stored data can be retrieved in several ways: (a) from documentation accompanying datasets, if it is provided; (b) from parameters' names; (c) by exploring dataset. The first option is time-saving and user-friendly as it does not require any additional steps, however, documentation is provided very rarely. The second option is widely spread, as it is a kind of "good practice" and nowadays often taken into account. For the presented example, the data publishers (Companies House of UK) provide documentation containing additional information about the published data. In addition, the names of almost all parameters are self-explaining and do not require

any additional analysis. However, this is the only such "user-friendly" dataset among four analysed Company registers.

According to (Kleppe et al., 2003), the PIM model is not related to the end-platform - it is independent of its specific and detail, and therefore does not include technical details. As a result, each parameter in a data object has name and an informal description of the value to be stored in it. Its notation is very simple and the corresponding data object with its 5 parameters is shown in Figure 4. But as part of this study, for each dataset, an in-depth analysis of data quality by analysing each attribute was performed. The extended (but not the full) data object is shown in Figure 5.

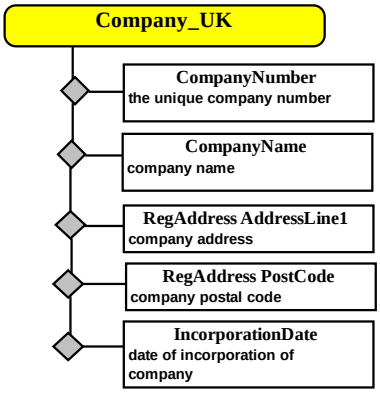

Fig. 4. PIM of data object "Company_UK"

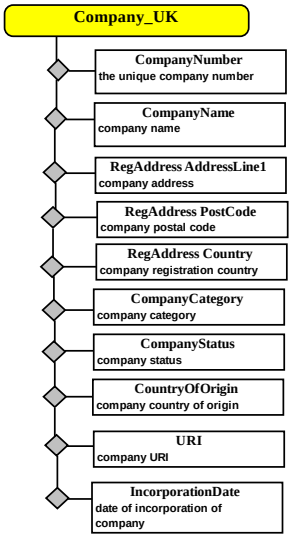

Fig. 5. PIM of the extended data object "Company_UK"

The proposed approach provides the possibility of modifying a data object as soon as the need arises, for example, when the use-case changes. This may be done by any person involved in the data quality analysis.

Compared to the PIM model, the PSM model must contain technical details. Descriptions of data objects' parameters are semi-formal at this stage as rules for attribute values syntax are provided. The syntax rules for describing the allowable values for the data object's fields can be formulated at different abstraction levels - from formal language grammar to definitions of variables in programming languages. In the latter case, the data object model is closely related to its implementation environment. The informal rules are replaced by formal rules at this stage, specifying more appropriate data type for each field depending on the values it stores. This information can be obtained: (a) from documentation about datasets provided by a data publisher; (b) from pre-processing, analysing data the most part of parameters contains. It should be noted that the format of the parameters also depends on the technique that will be used to replace non-formal descriptions with executable ones. The PSM model (Figure 7) is obtained from the PIM model depicted in Figure 5 (Figure 6 from Figure 4).

In the PSM model the corresponding data type is indicated for each parameter of the data object, indicating, if necessary, other characteristics of the parameter values based

on the data stored in the field (Nikiforova et al., 2020). Originally, the data type for all fields is a string (varchar). For example, it is specified that the name ("*CompanyName*") is an arbitrary length string, the value format of a date-containing field is "*DD/MM/YYYY*", the business form ("*CompanyCategory*") as well as the status ("*CompanyStatus*") is one of the acceptable values.

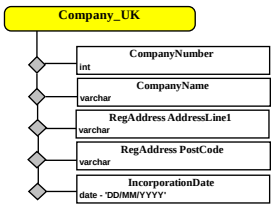

Fig. 6. PSM of data object "Company_UK"

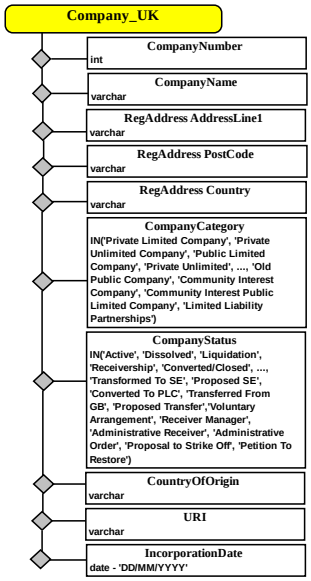

Fig. 7. PSM of the extended data object "Company_UK"

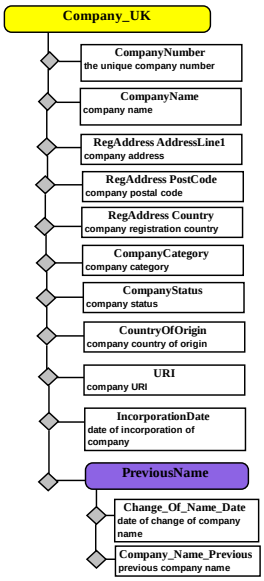

Fig. 8. Data object class "Company_UK"

Another concept that is required under a given solution is a "**data object class**". A collection of data objects of the same structure forms a data object class. The data objects class consists of several specific data objects, called instances, which are described by fields of arbitrary number and other data objects' classes. Each particular data object can have one to all parameter's values. It means the data object's class has a tree structure. In the above described particular example, Companies House collects not only current companies' names but also up to ten previous names. Two parameters describe previous name of company: the name itself "*Company_Name_Previous*" and the date when the name was changed "*Change_of_Name_Date*". Hence, the data object class "*Company_UK*" has 11 parameters, and the data object "*PreviousName*" has two parameters (Figure 8).

Data object class allows defining quality requirements for the data collection (Nikiforova, 2018a). It also allows to specify when quality is considered as high or low by introducing a threshold which cannot be exceeded. For instance, if the total error rate of quality problems of the data class "*Company_UK*" is lower than 5%, the dataset is considered to be of high quality, however, otherwise quality should be improved immediately. The total rank is calculated by relating the number of records having quality problems to the total number of records (corresponds with (Batini et al., 2016)). It also means that every user can introduce his own threshold that also goes in line with idea of the proposed solution.

4.3. Data quality specification

The second phase of data quality analysis is the definition of the quality specification for the data object defined in the previous step. A data quality specification for a specific data object consists of conditions that must be satisfied in order to consider the data object as of high quality (Nikiforova, 2019a). Data quality control of data object parameter values is reduced to the individual value check.

The data quality specification is retrieved from the data stored in specific fields or from the description of dataset. Usually data quality requirements may be: (a) retrieved from the data stored in specific fields or from the description of dataset, if the data provider provides it; (b) specified during pre-processing of the dataset or the subset; (c) since the quality of the data depends on the data user and the use-case, the requirements are defined by the user, i.e. they are based on the user's claims against specific dataset and data object. The first two options are only assistive, while the requirements are mainly defined by the user depending on the defined use-case. In the case of the Company House UK, it is partly retrieved from the documentation of data providers from which information on data, length, permissible values, etc. supplementing it with requirements which were formulated by users in accordance with their viewpoint. Some data quality requirements may be defined regardless of the use-case that meets objective requirements (Price et al., 2005). For instance, in the example of Company register, it is obvious that "*RegistrationNumber*" must have value that conforms to some pattern or corresponds to some specific format.

In the case of a PIM model, quality requirements are defined informally, for example by formulating them in a natural language or as formalised descriptions that are independent of implementation. They must be understandable to users who may not have in-depth knowledge of IT. The aim of this phase is to express clearly and comprehensively the requirements of the end-user that will be applied to the defined data object in the future. The quality requirements defined for each parameter are grouped together in accordance with the Mellor "classification" principle (Mellor et al., 2004). The PIM of data quality specification for the extended data object "*Company_UK*" is shown in the Figure 9, describing each condition in a natural language.

The 1^{st} – 4^{th} and 10^{th} boxes represent quality requirements for the data object depicted in Figure 6. Figure 9 demonstrates that, when checking the fulfilment of quality conditions for each parameter, error reports are prepared that contain values of the parameters do not comply with the required requirement. Error messages are recorded in the protocol for further processing at the data quality improvement stage. Regardless of the result of the prior verification, i.e. data compliance or non-compliance with the data quality conditions, a switch to the next parameter check until the last check is completed.

The next step is the transformation of the PIM model into the PSM. In the case of PSM, data quality requirements are replaced by formal requirements defined using logical expressions. The chart structure remains unchanged, changing only the quality requirement definitions by replacing informal ones with logical expressions. The names of the parameters of the data object serve as the operands of logical expressions, while both traditional programming languages - logical expression operations and data quality-specific operations can be used for operations. Logical expressions are both sufficiently expressive and easily understandable at the same time, which increases the possibility of users (without in-depth knowledge of IT and data quality) involvement in the process.

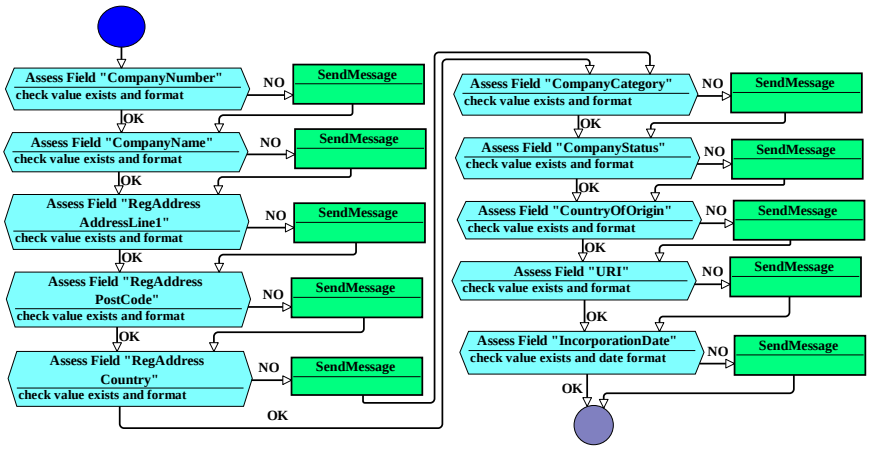

Fig. 9. PIM of data quality requirements specification

The nature of the logical expressions depends mainly on the PSM model of the data object and the quality specification PSM model defined in the previous stage. In the case of the Company House UK: (a) does the format of parameter "*IncorporationDate*" correspond to the defined? (b) is the value of the parameter "*CompanyCategory*" included into the list of allowable values? (c) does "*URI*" meet the pattern, in accordance with which every "*URI*" should start with a certain string while the second part should contain company's name taken from the first parameter ("*CompanyNumber*")? (see Figure 10).

Likewise to previous studies ((Nikiforova, 2018a, 2019a), (Nikiforova et al., 2019)), the most commonly used data quality requirements are: (1) existence of values, (2) relevance to specified type of data, (3) format of stored values (for example, length of the stored value), (4) conformity to a specific pattern, (5) relevance to the list of enumerable values, (6) validity of value (for example, trustful date) and other conditions that follow from the dataset and type of data that can be stored in the specific field. These requirements correspond mainly to declarative or objective requirements.

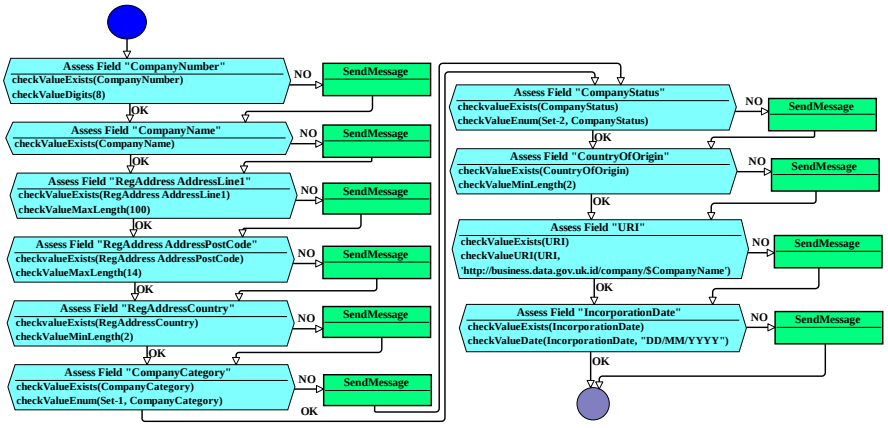

Fig. 10. PSM of data quality requirements specification

Data quality specification for the data object is described in a pseudocode written in the elements of the chart (Nikiforova et al., 2020). Despite the fact that pseudocode sometimes is related to PIM (for instance, in (Ruiz, 2018)), this time it can be considered as PSM since the pseudocode is closely related to the art of its implementation, for example, in programming language C# (conforms to ((Coutinho et al., 2012), (Kessler et al., 2010), (Shi et al., 2015) etc.).These requirements correspond mainly to syntactical checks, while contextual checks involving additional data objects are addressed in subsection 4.6.

4.4. Data quality evaluation process

Data quality evaluation process starts with description of activities that are necessary to be taken to select data object values from the data source. First, data objects' values are read from the data source and written into database. The complexity of this step depends on the data format, since the loading of structured data into the database usually does not cause any problems due to the similarity of their structure to the database table, while the selection and reading of semi-structured data may require additional actions that tend to depend on the structure of the document, for instance, data hierarchies, according to which, in some cases, a separate table should be provided for each level of hierarchy, linking them to each other through primary and foreign keys, as well as other features that depend on the data provider, such as incorrect selection of values and parameter separators (Nikiforova, 2018a). Then, one or more steps should be taken assessing data quality of the selected values, i.e. steps that should be taken when checking the data object's compliance with the data quality requirements defined. Data quality checks process data object classes. Data object instances are selected from the data source and recorded in the collection. All instances are cyclically processed, for each individual instance examining the fulfilment of the quality requirements, likewise in the case of processing an individual data object. The result of this process is the data quality problems identified for each individual instance. Therefore, if particular values don't meet defined data quality requirements, an appropriate message is sent (Figure 11). Non-empty "*SendMessage*" values form data quality problems' protocol that is saved in database for further processing. It can be used for improving data quality of particular dataset by triggering changes in the data source.

A PSM model of data quality is executable. The executability of the specification of data quality requirements allows the inclusion of quality requirements checks at different stages of data processing. This resolves data quality checking problems in situations where data are accumulated gradually, allowing a sequence of data entering a database to be different from that of entering "real world" or registering them. The data object class or the data object defined at the first stage is used as an input for quality evaluation process. Then, when data are read from data source and stored into a database, instances of the accumulated collection are inspected by verifying quality conditions, replacing quality conditions that were defined at the previous step by executable, for instance, SQL statements.

Figure 11 demonstrates the PSM of data quality evaluation process for the data object "*Company_UK*". The first element represents data reading and recording into a database for an operation followed by a parameter value quality check, automatically executing SQL queries.

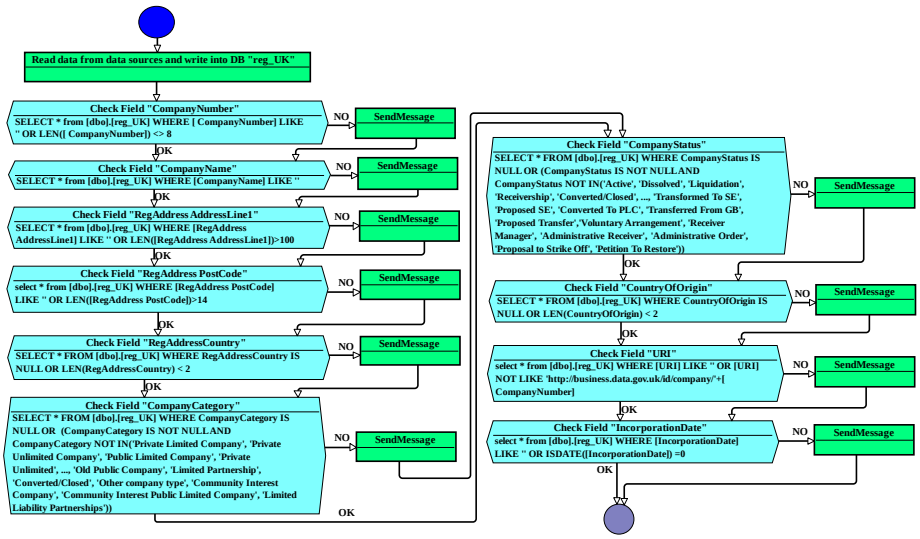

Fig. 11. PSM of data quality evaluation for data object "Company_UK"

SQL query design is close to the nature of data and data quality because the SQL query *SELECT* operator is suitable for data reading, while the *WHERE* condition is used to define data quality requirements by grouping the quality requirements defined for one parameter using *OR* operator. This option is also recognised by respondents as the most suitable for the given task, since the absolute majority of participants have preferred SQL language when defining executable requirements for the defined data object.

Nature of an executable artefact that replaces informal conditions depends on the person involved at this stage, his or her knowledge and experience since these conditions should be syntactically correct. This stage requires involvement of person with knowledge in IT area. It means, executable texts can be represented not only using SQL statements, they can be replaced by some programming language, for instance, C#. The implementation of this type of data quality requirement is close enough to data processing in relational databases, and the given model is platform independent because it is closely related to the execution environment. The main difference between the language describing the data quality evaluation process for a specific data object and the language describing the data quality requirements are data read from a data source or multiple sources and their record in the database. Considering the variety of data quality checks, depending on the use-case defined by the end-user, DSL has been defined in accordance with this principle, i.e. each time a data quality analysis is carried out, appropriate data quality checks are defined in accordance with a specific use-case. This phase is the first phase of data quality analysis, which could require the involvement of IT-experts.

Since the abovementioned mainly corresponds with syntactical control by checking the values of data objects within a single data object, i.e. the relevance of input data to their syntax, the need to extend the approach by providing the possibility of semantic or contextual verification within multiple data objects was identified. The description of the extension of the approach is described in section 4.6.

The last step is the execution of the PSM model. According to (Nikiforova et al., 2020), it can be implemented in several ways, two of which are: (a) the specification of quality requirements as programming work that allows precise formulation but is implemented with traditional programming methods; (b) a more general implementation option is an interpreter or compiler that would be able to execute the quality checks stored in the repository. The first option is usually found in IS where input data is entered by screen forms. The advantage of the proposed quality requirement specification is that it is possible, if necessary, to change the programme code according to the specification by separating the requirements specification from the source code. In object-oriented programming, it is possible to create a quality requirement check as a separate method that is applicable to a particular data object. Even though the second option is more complicated, the preference was given to it. For this purpose, the DIMOD tool was used as part of this solution. It shall ensure the creation of repositories during quality requirement modelling, allowing for changes in the models, without touching the programmes. A compiler is called for data quality checking, selecting the syntax used for quality requirements corresponding to a specific data quality analysis, C# or SQL, which is determined by the user involved in the process, based on his knowledge and skills in each technology. A data object is then transferred to the compiler and a description of the quality requirements is called from the repository. The compiler structure is based on linked lists where additional information can be recorded for each array element, i.e. a reference to the associated list is stored. A chart that is defined in the DIMOD tool is converted to a graph by passing on which it is converted to an executable code. According to the abovementioned, the stage of the data quality checking process can suppose the involvement of IT-specialists, which is caused by one of the main limitations - correctness of the formal executable texts defined in the chart of all data quality checks, as the DIMOD tool and compiler do not perform checks on the validity of inputted text, i.e. this must be ensured by "smart users" (Bicevskis et al., 2018a).

A brief description of the proposed approach implementation is given in the next subsection.

4.5. Implementation

According to abovementioned and (Nikiforova et al., 2020), the presented data object-driven approach uses graphical DSLs for defining data quality models. Every component of data quality model – data object, data quality requirements, data quality evaluation or assessment process - are described using its own graphical DSL.

Since it is highly recommended to not use one specific graphical editor that supports only one DSL since it can become very complex, a tool-building platform DIMOD was used in this study, which allows to define many different DSL with different data-object structures. DIMOD is a derivative of the graphical tool building platform GrTp, developed in LUMII (Barzdins et al., 2007). It should be noted that the GrTp solution from which the used DIMOD tool was derived is also based on MDA principles. In scope of the research, a three-language family was created through DIMOD, the definition and representation of which is provided by it. Each DSL language has its own structure that corresponds to the examples viewed in the previous subsections. Due to its DSL configuration capabilities (Sprogis et al., 2013), when a DSL metamodel is entered to the repository, DIMOD converts it to a graphical editor, which, when interprets the DSL metamodel, offers all the capabilities needed for a graphical editor - drawing

graphical charts, editing them, creating tree structure models, and other actions. The advantage of the solution is that models are automatically saved as tools defining metamodel instances that is achieved by the Configurator tool. It is a DSML tool that allows to specify tools at a higher level of abstraction compared to UML class charts. For interpretation of specifications, or processing with a universal interpreter, the corresponding specifications are automatically transformed into a universal metamodel. As a result, the users, or developers, do not need to be familiar with the tool definition meta-models, its default values, or to make sure the created models are correct. This significantly reduces the development time of the tool definition metamodel instance, eliminating the possibility of building an interpretation that is not appropriate for the interpreter.

According to (Kleppe, 2008), the central component of the DSL definition is abstract syntax, which plays a primary role in the language specification. It defines language concepts and their relationship, including limitations on model creation. The definition of abstract syntax was performed by means of a metamodeling technique which according to (Akehurst et al., 2002) is considered to be a better form of normalization of [graphical projects], which guarantees the integrity of the data in the model through formal techniques, reducing and even eliminating redundancy and the possibility of various types of anomalies. Metamodels are sufficiently expressive and easy to understand, particularly when compared to textured syntaxis, beforehand the use of graphic grammar. For this reason, metamodels are preferred, when complex relationships between language concepts take place that cannot be simply described through textual syntax. In addition, metamodeling allows to combine different constraints that would be stored separately in case of context-free grammar (Selic, 2009). In general, the use of metamodels significantly simplify the definition of a language, allowing, for instance, (a) a graphical representation of each element, (b) depicting "*from*" and "*to*" relationships using arrows, etc., that in the case of text-based syntax, would be difficult to explain since many additional steps need to be taken to define each element (retaining the link to its shape, colour, etc.). In this case, these elements are defined in a more convenient and easy-to-replicate way.

Creating a metamodel is one of the most difficult steps that requires appropriate modelling knowledge. When the configured graphical editor is prepared, it can be published in WEB for its further use, therefore users can take advantage of the opportunities provided without thinking about creating metamodels, spending their time on creating of appropriate graphical charts only, the structure of which is intuitive and close to the nature of data and data quality. This means that once an editor is configured, it becomes reusable. It should be noted that not only graphic editors can be published online, but also charts already created that allow end-users to explore pre-created charts before they form their own charts. Providing this option can serve as a kind of guide for end-users.

4.6. Contextual data quality analysis

In practice, it is often not enough to check data quality within a single data object, requiring contextual data quality analysis within multiple data objects. Contextual or semantic control is characterised by verifying that a data object is appropriately related to other data objects or is compatible with other values of data objects already entered in the data source, determining whether the data is inconsistent. The nature of semantic

control requires repeated semantic control each time once the values of interrelated data object attributes changes.

Traditionally, a semantic or contextual check is carried out in two steps: (1) find the corresponding entry in an "external" dataset; (2) validation of the fields' record of an initial dataset against a found record (Scannapieco et al., 2005). In (Batini et al., 2016) the first step is called "*record identification*", while the second is "*decision strategy*". This means that all matching records in both datasets are initially found, linking datasets to specific parameters, followed by a cross-compliance check for the values of each matching pair. In this case, all relevant parameter values of the primary data object form a subset of all values of the corresponding parameter of the secondary data object. The two sets may fully overlap, but the matching parameter set of the primary data object must not contain elements that are not available in the secondary data object. Of course, this means that checking the values of a particular parameter requires high quality and completeness of a secondary data object. (Batini et al., 2016) recommends making a decision at the "*decision strategy*" stage or, if the values coincide, it is possible to claim that both values represent the same real-world object. In other studies, value matching is enough to admit that equal values point to the same object. Taking into account the objective of the proposed approach - to customize quality analysis to the end-user as much as possible, in this case, the user shall decide whether value matching is sufficient, or at the same time other conditions must be fulfilled, for example by comparing the values of other parameters.

According to (Nikiforova and Bicevskis, 2019), the need for a contextual data quality check is observed in the case of the Companies House UK, which requires a contextual check of the values of [*CountryOfOrigin*] and [*RegAddress Country*] containing countries names (Nikiforova et al., 2019). The analysis of data quality within a single data object does not allow for an unambiguous decision on the quality of the values in question, resulting in only potentially poor records and values. In order to take a decision on their quality, it is necessary to compare values to countries names that meet the standards, i.e. data objects where existing data must be of high quality. For this purpose, a "*Country*" data object whose parameters *ISO*, *ISO2*, *ISO3*, *UNI*, *UNDP* meeting a certain standard of national names (FAO, 2019) was created.

The involvement of additional data objects requires the division of "**data object**" concept into **primary** and **secondary data objects**. The **primary data object** is considered to be a data object whose quality is analysed, which is the central object of data quality analysis. A data object is considered to be a **secondary data object** if it forms the context of an analysed or primary data object.

Both primary and secondary data objects are defined by the end-user, whereby all the principles and characteristics of the primary data object are also applicable to the secondary data object. The number of secondary data objects involved in the data quality analysis depends on the nature and the use-case defined by end-user. Similarly, their number is determined by the nature of the primary data object and their parameters - how many secondary objects can be defined and linked to parameters etc. if any. The primary data object is typically one - the central object of the data quality analysis, the quality of which is of interest to the end-user, which may be associated with an unlimited but final number of secondary data objects (Nikiforova and Bicevskis 2019).

A secondary data object may consist of (a) another dataset independent from the primary data object, (b) a data object retrieved from the primary data object that may be used to validate the permissible values. The second option is intended to simplify the

corresponding data quality requirement by excluding the possibility of including all allowable values in the SQL query, significantly expanding it. The definition of a secondary object filled with all allowable values also ensures their re-use in other data quality analyses.

As for the example examined, as in the case of one data object, a number of different data quality requirements may be defined for the analysis of the primary data object against the secondary data object, for example:

1) the country name of the register must comply with at least one country name from the standard. This makes it possible to make sure that the existing values of the Companies House UK are valid - compatible with the real world;
2) all existing countries names in the register must meet the same standard within the same dataset. This allows to make sure that the data of parameters specified in the dataset are homogenous. Two options are also available, depending on the end-user:
    2.1) conforms to one of the generally accepted standards;
    2.2) comply with a generally accepted standard specified by the user.

A contextual data quality test was performed in line with requirements 1 and 2.1, although the concept of data quality in its traditional sense does not comply with option 1, i.e. a value validity check.

The given quality contextual check requires modifications to the charts available in the preceding sub-sections according to the given task, following classical contextual data quality evaluation principles (Scannapieco et al., 2005). In the phase of a data object definition, when checking the quality of an analysed dataset against another dataset, a secondary data object(-s) is defined in addition to the primary data object. The example provided contains only one secondary data object, but (Nikiforova, 2019) shows a primary analysis of the quality of the data object against 3 secondary data objects. In terms of graphical representation, the secondary object definition corresponds to the primary data object, assigning a different colour to the secondary data object. The relationship between the primary and secondary data object(-s) is represented by arrows. The relationship between multiple data objects, indicating which parameter of a secondary data object is associated with a particular parameter of the primary data object is demonstrated. A more detailed relationship between data object parameters is defined by the quality requirements in the next phase, i.e. the definition of the data quality specification, which defines the requirements against the quality of fields in a primary data object in the context of a secondary data object.

All the conditions defined in the following steps focus mainly on the primary data object: the proposed solution does not analyse the quality of the secondary data object, as it is a supplementary for analysis of the data quality of the primary data object. Considering the importance of the data quality of a secondary data object for the analysis of a primary data object, it is intended that its quality was checked in advance, defined as a primary data object, or considered to be of sufficient quality.

Figure 12 defines the primary data object "*Company_UK*" and the secondary data object "*Country*". The corresponding parameters of the primary data object [*RegAddress Country*] and [*CountryOfOrigin*] are linked by arrows to the relevant parameters of the secondary data object "*Country*". According to the pre-defined first use-case, the primary parameter value must correspond to at least one of the parameter values of the secondary

data object. Both parameters of the analysed primary data object are associated with each parameter of the secondary data object. Depending on the use-case, each parameter in the primary data object could be associated with different parameters of the secondary data object, or none of them.

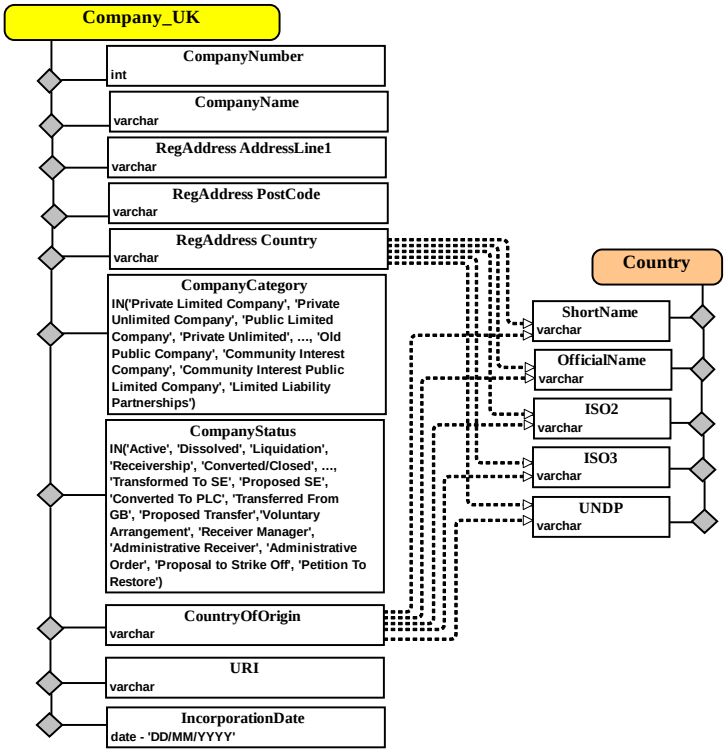

Fig. 12. Data object "Company_UK"

In the case of the second use-case, the value of the parameter of the primary data object must correspond to a generally accepted standard established by the users, each parameter of the primary data object would be associated with one parameter of the secondary data object. It should be noted that when a use-case changes, the corresponding chart could be adapted to it at any stage of the analysis. When a data quality analysis should be done based on a number of different use-cases, charts can be re-used without requiring the user to re-define it since when it is created for the first time, it was stored for its further use.

As for a data quality requirement specification chart in case of contextual verification, the chart described in section 4.3 is supplemented with data quality requirements for checking the relevant parameters of the primary data object against the parameters of secondary data objects. The nature of the requirements corresponds to the requirements for the quality analysis of single data object. When defining the quality requirements for the parameter values of the primary data object in the context of the secondary data object(-s), some data quality requirements within one data object may not only be supplemented but also replaced by others. For example, in the case of the parameters analysed, pre-defined requirements "*CheckVailueExists*" and "*CheckValueMinLength*

(2)" are replaced by checks against the values of the parameters of the secondary data object, so the initial checks become redundant, there is no need to verify the existence and length of a value because the correctness of values against values of the secondary data object are checked (see Figure 13).

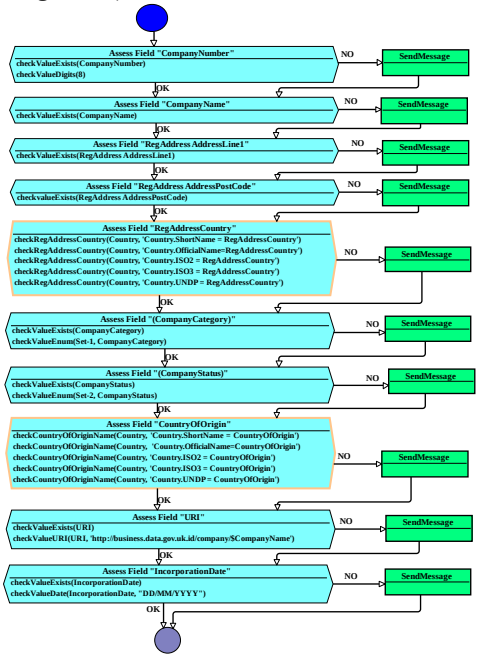

Fig. 13. Data quality specification for data object "Company_UK"

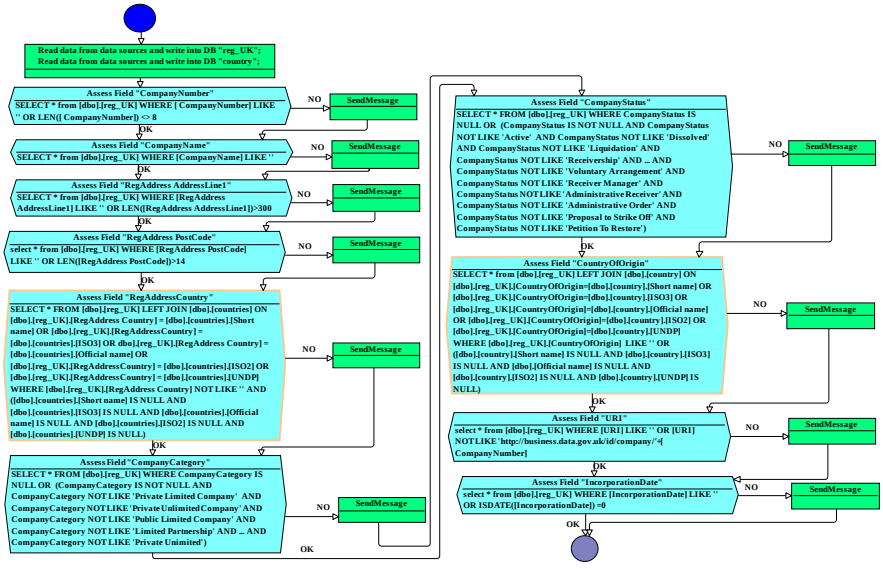

Fig. 14. Data quality evaluation for data object "Company_UK"

At the data quality evaluation stage, data quality requirements defined by logical expressions are replaced by executable data quality requirements (Figure 14). The corresponding changes to the chart are: (1) data reading operation from a secondary data source is added, (2) data quality check for the relevant parameters are defined.

Thus, the proposed model has been extended to contextual data quality control by checking the data quality of the primary data object against an unlimited number of secondary data objects. This allows a deeper analysis of data quality by investing less resources in it.

Contextual analysis was performed for 18 datasets. Contextual data quality problems were identified in 17 of them (94.4%), without identifying quality problems in the 18th dataset, however only after matching data samples for their future analysis, which can be also considered to be a data quality problem. In terms of the number of parameters, contextual analysis was performed for 61 parameters, with at least some data quality problems identified in the 83.6% parameters. Since the identified problems were not identified as a result of data quality analyses within a single data object, only through individual checks, identifying potential problems, failing to make a decision on their non-compliance with the real world, there is a reason to argue that the possibility of contextual checks makes it possible to significantly improve data quality results by providing an opportunity for deeper and more comprehensive analysis of data quality. The detail and the nature of the problems identified are discussed in (Nikiforova and Bicevskis, 2019).

4.7. Summary

The proposed data object-driven approach to data quality evaluation vitally differs from the existing ones. Its idea is not presented in other works, as demonstrated by the analysis of existing solutions as well as by Batini, the world's leading researcher, author of a deep study of the data quality problem and an overview of existing methodologies, published in a number of books and scientific articles (Batini et al., 2006, 2009, 2016). At the same time, it should be noted that the proposed solution is simple and even intuitive, because it is close to the nature of data and data quality.

The proposed solution is an "external" mechanism that allows data quality analysis to be carried out by data users without knowing how data was accrued and processed with data providers. This means that it: (a) can be applied to "third party" data - both "closed" and "open" data; (b) is intended for both data providers and data users. It should be noted that the approach is intended to analyse the quality of both structured and semi-structured data.

Considering the simplicity of basic concepts of the proposed approach, their unambiguous and clear definition, the use of graphical DSL and the involvement of two models, it is possible to assume that the approach is appropriate for users without in-depth knowledge of IT and data quality, since the involvement of IT-specialists can only be needed at the final stages, by replacing informal descriptions with executable ones. The proposed approach supports and even facilitates users' cooperation, allowing a number of individuals to be involved at any stage, if necessary. In addition, any type of change may be initiated as soon as it becomes necessary, i.e. at any stage of data quality analysis. Therefore, the "flexibility" of the solution has been achieved.

The definition of data object and data quality specifications are fully dependent on the particular data user and his/ her use-case, which ensures that the data quality analysis is accurate in line with the expectations and needs of the data user.

It should be noted that the potential of the proposed solution is not limited to the quality analysis of the individual datasets, since (Bicevskis et al., 2019b) describes the data quality runtime verification that is based on the proposed data quality model. This allows to verify the quality of the data in almost real time by performing a data quality analysis during the execution of the business process. This means that it is possible to verify that a particular process did not damage the data in the system, but if any data quality requirements were violated during the execution of a business process, this business process is identified in the same way as non-qualitative data as in the case of a quality analysis of the dataset. This makes it possible to ensure continuous analysis of data quality.

In addition (Nikiforova and Bicevskis, 2020) demonstrated the potential of the proposed data quality model by using it as a test model for data quality model-based testing (DQMBT).

Moreover, (Bicevskis et al., 2019a) presents the formalisation of the presented solution, with the aim of transforming it into a data quality theory which despite the high number of data quality solutions developed in recent decades, hasn't been proposed yet. One and probably the main reason why data quality theory has not been proposed up to now, despite the large number of attempts, is that any theory must be based on clearly defined concepts. Considering that data quality is traditionally associated with the concept of data quality dimension, which lacks a universal definition, classification, and measurement mechanisms, this vital requirement cannot be met. The proposed approach does not use the concept of data quality dimensions, using the more general term *"data quality requirement"*, otherwise following up all commonly accepted definitions of data quality and related concepts, which implies that clearly and unambiguously defined components of the proposed data quality model and the specific nature of the proposed solution make it possible to offer an informal data quality theory through the formalization of this approach.

However, the proposed solution also has its own limitations. First, the current solution does not include verification of the validity of text recorded in the model elements, which according to the abovementioned should be ensured by "smart users" (Bicevskis et al., 2018a). This restriction is based on the limitations of the DIMOD tool and the compiler developed.

Secondly, the solution is intended to analyse the quality of structured and semi-structured data, therefore it is not intended to analyse the quality of unstructured data. As the solution is intended for users without in-depth knowledge in the fields of IT and data quality, the solution also needed to be as simple as possible, while adapting the solution to the analysis of unstructured data would make it more complex. However, it should be noted that in some cases it can also be applied to unstructured data. For example, when the text should be processed, and the end-user wants to make sure that facts in it are semantically correct, it is possible to define a data object that will contain the attributes whose values will be analysed by storing the corresponding values in it, and their further analysis would be performed according to the previously discussed procedure. This would work well in the case of the processing of a text with a number of uniform objects and their description, such as descriptions of different countries, including their name, capital, area, official language, etc.

Thirdly, the solution is mainly suitable for data quality analysis without focusing on the analysis of datasets' compliance to the open data principles, such as metadata analysis. However, despite the fact that the solution was not originally intended for

metadata quality analysis, it can be used for metadata analysis as well, by considering metadata describing specific datasets as a data object, and defining data quality requirements for this data object, the quality of metadata may be evaluated according to a previously discussed procedure.

5. Results of application of the approach

The proposed data quality assessment approach was applied to several datasets (some of them have been published in (Nikiforova, 2019, 2018a, 2018b), (Nikiforova et al., 2019), (Bicevskis et al., 2018b)), summarizing (1) the results of the quality analysis of company registers of the UK, Latvia, Estonia and Norway, (2) a brief description of the experience gained in datasets, highlighting the most common open data quality issues, (3) focusing on datasets representing one specific domain and analysing their quality - open medical/ health[care] data from Latvia. It should be noted that the datasets analysed are open datasets provided by different data providers/ publishers, so that the results of the analysis allow an assessment of the overall quality of the open data, the degree of which cannot be attributed to the data provider.

It should be noted that this approach can be applied not only to open data, but also to a variety of structured and semi-structured data, but open data analysis allows (a) to verify the quality of data freely available to users by assessing their usability, (b) apply the proposed access to data without violating privacy, security, and privilege restrictions, while showing that such approach may be applied to "foreign" or "third-party" data without knowing how they were collected and processed.

A comparative data quality analysis of four European Company Registers (Latvia, Estonia, Norway and the United Kingdom) is available in (Bicevskis et al., 2018b), while the demonstration of contextual analysis results on the example of the Company House UK has been published (Nikiforova et al., 2019). The first stage of data quality analysis of Company Registers was based on two use cases defined for all Company Registers by comparing them, however, at the second stage of analysis data quality of each Company Register was analysed only within one Company Register, analysing each company characteristic parameter by performing in-depth analysis of registers.

Since the defined use cases are simple enough, where only the quality of primary attributes is analysed, it was assumed that the data should be (a) complete, (b) free of dubious values, (c) correct. However, the results of the data quality analysis have shown that this assumption is incorrect.

According to the first use case, the possibility to find/ identify every *Company* by its name, registration number and date of incorporation was examined. The results of the analysis summarized in Table 1 which shows that Company Registers of the UK and Latvian have records of companies that do not have a name, as well as data quality problems in the date of incorporation. No problems were found in the Estonian and Norwegian Company Registers, but it should be noted that the Company Register of Estonia does not provide data on the date of incorporation of the companies, so it does not correspond fully to the use case. 12 incorporation dates of Companies are dubious, since one of Norwegian companies was registered in "*1277-09-13*" and one of the UK companies - in "*25/04/1552*", that is unlikely.

As for the second use case, simple quality checks of the address and postal code values were performed aimed to check (a) the existence of an address value, (b) the existence of a postal code, and (c) its correspondence to a particular sample, defining its

own format or pattern that depends on country. Several data quality problems were defined in all Company Registers. As for address parameter, the best results were demonstrated by Company Register of Latvia (0.09% quality defects), followed by the United Kingdom (0.997%), Norway (6.2%) and the worst - Estonia (11.24%). In the case of postal codes, Company Registers of Norway and UK have the lowest number of data quality problems (1.3% and 1.6%), followed by Latvia (5.16%), while the highest number of data quality problems is found in the Company Register of Estonia (8.5%).

Table 1. Summary of the results of the analysis of the quality of Company Registers.

Company Register	Name	Registration number	Incorporation date	Address	Postal code
UK	1 (0.0001%)	0	3 (0.0004%)	7 518 (0.997%)	12 151 (1.6%)
Latvia	10 (0.0025%)	0	94 (0.02%)	366 (0.09%)	20 498 (5.16%)
Estonia	0	0	-	29 918 (11.24%)	22 621 (8.5%)
Norway	0	0	9 (0.0008%)	68 128 (6.2%)	14 683 (1.3%)

This means that none of the very simple or intuitive and even obvious use cases in which the values of the primary parameters were analysed were satisfied by any Company Register. However, the Estonian and Norwegian Registers can be used to identify any company by its name and registration number, since only they have passed quality checks of the relevant fields. However, the existence of data quality problems in other fields does not indicate that the corresponding datasets are of low quality and cannot be used by users, as the number of data quality problems detected is relatively small and could be quickly improved, for example, using the proposed approach. These enhancements are not too resource-consuming, however, would significantly improve the overall quality of the data.

It is important to note that the existence of data quality problems in datasets is crucial because data providers are not even aware of them. This also corresponds with *Global Open Data Index* (WEB, a), which evaluates 15 public sector open datasets, including Company Registers, where each Company is represented with its name, unique identifier or registration number and address, that fully corresponds with primary attributes used in previously described analysis. As a result of this analysis, the *Global Open Data Index* has ranked Company Registers of Norway and the United Kingdom at the 1st position, and Latvia – 18th out of 94 Company Registers. Such high results are explained by the fact that the *Global Open Data Index* evaluates the compliance of specific datasets with the open data principles without evaluating data quality, which corresponds with the claim of ignoring the data quality dimension in the open data principles list. This also means that end users should be aware and not rely on such estimates, since high scores do not necessarily indicate high data quality of datasets.

The quality of the analysed datasets was also checked for other parameters characterizing the data object (i.e. "*Company*"). As a result, several data quality problems were identified in each Company Register, both in data syntax and semantics. (Nikiforova, 2018a, 2019) summarize the results of the data quality analysis of each Company Register, providing a discussion of the identified data quality problems, their nature and feasibility.

5.1. Results of data quality analysis of Latvian open health[care] data

Considering the importance of medical data, the analysis of Latvian open health[care] data quality was performed. Considering that data quality problems in Latvian health data are usually found in "closed" data, mainly referring to data inconsistency, there are reasons to believe that data quality problems will be detected in open datasets. Open health data was first published in Latvia in 2018 and in the third quarter of 2019 they were represented with 15 datasets published by 7 different data providers.

Table 2. The most common data quality problems by data set.

Dataset	Context issues/ context total	Empty/ Total	Multiple notation/ Total	Defects in interrelated parameters (yes/ no)	Clean / Total
Incidence of 2nd type diabetes in Latvia	0/0	0/6	0/6 (0)	no	6/6 100%
Distribution of persons receiving tech aid by AT	2/2 (100%)	3/7 (43%)	0/7 (0)	no	2/7 29%
Number of social service providers	2/2 (100%)	22/27 (82%)	10/27 (37%)	no	4/27 15%
Persons with disabilities by the severity of the disability and AT	2/2 (100%)	0/23 (0)	0/23 (0)	no	20/23 87%
Number of children with disabilities by AT	2/2 (100%)	0/10 (0)	0/10 (0)	no	8/10 80%
Accidents at work	(0-1/1) (0-100%)	1/10 (10%)	0/10 (0)	no	8/10 80%
Occupational diseases confirmed	4/5 (80%)	2/11 (18%)	1/11 (0.09%)	no	9/11 82%
National Blood Donor Center Statistics	0/0	0/4 (0)	0/4 (0)	no	4/4 100%
Register of licensed pharmaceutical companies	1/2 (50%)	17/38 (45%)	0/38 (0)	no	19/38 50%
Medicines consumption statistics	3/3 (100%)	5/8 (63%)	2/8 (25%)	no	0/8 0
Medicinal Product Register of Latvia	4/9 (44%)	21/41 (51%)	1/41 (2%)	yes	14/41 34%
Food supplements register	2/2 (100%)	30/35 (86%)	4/35 (11%)	yes	5/35 14%
Dietary foodstuffs register	2/2 (100%)	19/22 (87%)	4/22 (18%)	yes	3/22 14%
Veterinary medicinal product register	1/3 (33%)	16/26 (62%)	0/26 (0)	yes	8/26 31%

As the focus of the study is data quality, despite the ability to analyse only the quality of individual parameters in scope of one particular use case, each parameter in each dataset was analysed with the aim of performing in-depth data quality analysis. 15

primary datasets and 11 secondary datasets were used for data quality analysis. The most common data quality problems in the analysed datasets are: (a) contextual data quality problems; (b) incompleteness of the data; (c) different notation of one object within one data object and even within one parameter; (d) data quality issues in the case of interlinked parameters.

Table 2 provides a summary of the most popular data quality problems by the number of parameters in which they were found, while (Nikiforova, 2020c) summarizes the number of records in which the particular data quality problem was identified as well. Only one dataset does not have any data quality issues (last column of Table 5.2.1). It should be noted that it collects numerical data, which also made quality checks simpler, mainly through completeness checks and simpler mathematical calculations related to data aggregation through data quality analysis within a single dataset. A detailed overview of the most common problems, with appropriate examples, is available in Section 5.3.

5.2. Summary of the results of the data quality analysis of the datasets

An analysis of data quality of open datasets within the study concluded that 83.3% of the datasets had at least some data quality defects, however, neither the data users who are free to use the data for their own purposes in processing, analysing and using them in decision-making, nor the data providers who have published and used the data, [probably] are aware of their existence in their information systems.

The most common data quality problems are:

- data incompleteness, even in primary data (77% of datasets analysed);
- contextual data quality issues identified in the 83.6% parameters;
- different notation of one object within one data object and even one parameter, or inconsistency of values and different values to denote one real data object;
- data quality issues in interrelated parameters.

In scope of the given research data quality defects in data quality data analysis within one dataset were detected in 83.3% cases. As a result of contextual analysis of data quality in scope of several datasets, data quality problems were identified in 94.4% of analysed datasets. This means that open data have data quality issues.

Data quality issues in open data appears even when very simple use-cases are chosen. The results of the analysis demonstrate that several identified data quality problems are systematic and can be observed in data of specific domains, in both cases, i.e. of national and international datasets. Most data quality issues can be resolved by making minor changes that are not too resource-consuming (both in terms of time and human resources) if they are systematic or have a specific value that is common to multiple records, i.e. even correcting one value could significantly improve the overall quality of the dataset while simultaneously solving problems across multiple records.

One of the main problems in using open data is that the data quality data is not known to end users, and even more, it is not known under which requirements or use-case they will be qualitative enough and useful for analysis and decision-making and when the use of data will lead to inaccurate or even invalid results. Determination of the suitability of a specific dataset for the needs of the user and his use case, the quality of the dataset must

be pre-tested, making sure that it satisfies the conditions of the particular use case. This can be achieved through the above procedure.

Taking into account that in scope of this research all analysed datasets were analysed in-depth, approaching "*absolute*" data quality, testing datasets for multiple (but of course not all) possible uses, created diagrams and tests are more complicated than those that will be defined by end users analysing data quality for their own purposes. Taking into account that all components of the proposed approach are simple, unambiguous, and intuitive, as is the entire quality model, there is reason to believe that it is appropriate and will be used by a wide range of users, including users without advanced knowledge of IT and data quality. This would provide not only data quality testing for user's own purposes, but would also facilitate collaboration with data providers, thus contributing to data quality improvement towards absolute data quality.

6. Conclusions

The paper addresses [open] data quality. In scope of the study, literature on data quality issue, its relevance, existing approaches to data quality analysis and evaluation was studied, and an alternative data object-driven approach was proposed.

To sum up, "data quality" is a complex concept of a relative nature, according to which data quality is the suitability of the data for the use-case of a particular user. The quality of the data depends on the context, and as the data in IS changes over time due the gradual accumulation, the data quality requirements may change over time. Despite this challenge is old, that is proved by early studies, the popularity and the topicality of the data quality problem are also growing rapidly, mainly due the open data and their popularity. However, the quality of open data is unreasonably little researched, despite new challenges arising from their nature (i.e. the ratio of open data quality studies against the total number of data quality studies, is ~ 0.2%).

The study carried out an analysis of more than 70 existing solutions, concluding that existing studies are mainly (a) general studies on data and information quality, mostly focusing on definition of data quality dimensions and their groupings; (b) quality assessment of open data portals and/ or Open Government Data; (c) quality assessment of linked data. The analysis of existing solutions demonstrates that the majority of studies are not suitable for users without in-depth knowledge in the fields of IT and data quality, which are not acceptable in the current circumstances, since the majority of users have daily contact with data and they should be able to check their quality. The need for such possibility is also related to the popularity of the open data. In addition, most existing solutions use a high number of data quality dimensions and require the definition of data quality dimensions and requirements, as well as their application to appropriate defined or pre-defined dimensions which tend to cause difficulties even for data quality professionals. The majority of existing solutions require the involvement of data quality- and IT- experts at all stages of data quality analysis. However, the analysis of data quality depends on the use-case as well as open data can be used by any end-user, therefore the end-user must be involved at all stages of data quality analysis, reducing the involvement of IT-experts.

In order to resolve the problems identified, a vitally new data object-driven approach to data quality analysis was proposed. This can be described as follows:

1) the proposed data quality model consists of 3 main components: (1) data object which quality is analysed; (2) data quality specification which depend

on the use-case; (3) data quality measuring process. The proposed approach does not link the quality of the data to the concept of "data quality dimension" by replacing it with a more universal concept of "data quality requirement", which is a superset of the data quality dimensions related to the quality of the data;

2) the specification of data quality requirements (data quality model) that is defined in DSL concepts, is executable. The quality assessment process results a protocol containing non-conformities with the quality specification found during the inspection process, that may be used to improve data quality;
3) the proposed approach is the user-oriented approach, where each component of the proposed data quality model is defined by the end-user by verifying the quality of the data of a specific dataset for its own purposes, focusing only on those parameters describing data object that are important within a specific analysis. Thus, the results of the data quality analysis are as close as possible to the original intentions of the end-user, i.e. to the understanding of the concept of data quality;
4) each component of the proposed approach is defined using a relatively simple DSL language. Created models are easy to create, edit and reuse. The proposed DSL platform provides a wide range for the definition of languages of quality requirements, allowing modular definition of quality requirements and the verification of the requirements at different stages of data processing, without including data quality requirements in one comprehensive requirement specification, where data quality is checked only at the end of data collection;
5) most steps in data quality analysis do not require users to have prior in-depth knowledge of IT or data quality. The data quality analysis process is becoming intuitive, which makes it possible to assume that the approach is intended for a broad range of users. The involvement of IT specialists may only become necessary at the final stage, transforming informal requirements into executable, so that IT-specialists carry out a support function without affecting the definition of the basic components of data quality analysis, i.e. the data object and the data quality requirements applicable to it;
6) as a result of the application of data object-driven approach to data quality evaluation to the real "open data", (a) the appropriateness of the proposed approach to "third-party" data quality analysis was demonstrated, (b) a number of different types of data quality problems that are typical of both Latvian and foreign open data were identified. When performing data quality analysis for datasets within a single dataset, data quality problems were identified in 83.3% of analysed datasets. As for special domain - Latvian open health[care] data, a number of data quality problems were detected that were classified in several categories according to their nature. As a result, it can be concluded that certain types of data quality problem may be considered to be a widespread trend;
7) an extension of the proposed approach ensuring a possibility to perform data quality analysis in scope of several data objects that may be obtained from different data holders and, thus, performing in-depth quality analysis leads to a significant improvement in the results of applying the approach. This is

reflected in the results of its application to 18 datasets, data quality problems identifying in 17 of them (94.4%). In terms of the number of parameters, contextual analysis was performed for 61 parameters, with at least some data quality problems identified in the 83.6% parameters. It significantly improves the results of data quality analysis, as well as require fewer resources to carry out them, as well as ensuring data objects reusability;

8) despite the existence of data quality problems in 83.3% analysed datasets, most of them could be resolved by making small revisions that do not require a lot of resources, since the greatest effort requires their identification, which can be relatively easily achieved through the proposed approach.

The developed approach is an external solution that allows the analysis of the quality of "third-party" datasets, regardless of the system in which data was stored without requiring knowledge of their storage and processing mechanisms.

The wide scope and relative simplicity of the proposed approach make it possible to assume that the proposed approach will be used not only by end-users for their own needs, but also with open data quality enthusiasts who are becoming popular around the world, including Latvia. Involvement of enthusiasts in the analysis of open data quality and the use of feedback would contribute to the overall improvement of data quality by approaching "*absolute*" data quality – data quality that satisfies all possible use-cases.

Acknowledgments

This work has been supported by University of Latvia project AAP2016/B032 "Innovative information technologies".

References

Aarshi, S., Malik, B. H., Habib, F., Ashfaq, K., Saleem, I., Tariq, U. (2018). *Dimensions of open government data web portals: A case of Asian countries*. International Journal of Advanced Computer Science and Applications, 9(6), 459-469.

Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., Lehmann, J. (2013). *Crowdsourcing linked data quality assessment*. In International semantic web conference (pp. 260-276). Springer, Berlin, Heidelberg.

Akehurst, D. H., Bordbar, B., Rodgers, P., Dalgliesh, N. T. G. (2002). *Automatic normalisation via metamodelling*. In ASE 2002 Workshop on Declarative Meta Programming to Support Software Development.

Attard, J., Orlandi, F., Scerri, S., Auer, S. (2015). *A systematic review of open government data initiatives. Government Information Quarterly*, 32(4), 399-418.

Barzdins, J., Zarins, A., Cerans, K., Kalnins, A., Rencis, E., Lace, L., Liepins, R., Sprogis, A. (2007). *GrTP: transformation based graphical tool building platform*. In The 10th International Conference on Model-Driven Engineering Languages and Systems, Models.

Batini, C., Scannapieco, M. (2016). *Data and information quality*. Cham, Switzerland: Springer International Publishing. Google Scholar.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A. (2009). *Methodologies for data quality assessment and improvement*. ACM computing surveys (CSUR), 41(3), 16.

Batini, C., Pernici, B. (2006). *Data Quality Management and Evolution of Information Systems*. In IFIP World Computer Congress, TC 8 (pp. 51-62). Springer, Boston, MA.

Bauer, F., Kaltenböck, M. (2011). *Linked Open Data: The Essentials, edition mono/monochrom*. Vienna, Austria.

Bevan, C., Strother, D. (2012). *Best practices for evaluating method validity, data quality and study reliability of toxicity studies for chemical hazard risk assessments*. Washington (DC): American Chemical Council, Centre for Advancing Risk Assessment Science and Policy.

Bicevskis, J., Nikiforova, A., Bicevska, Z., Oditis, I., Karnitis, G. (2019a). *A step towards a data quality theory*. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE.

Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2019b). *Towards Data Quality Runtime Verification*. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE

Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2018a). *An Approach to Data Quality Evaluation*. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 196-201). IEEE.

Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2018b). *Data quality evaluation: a comparative analysis of company registers' open data in four European countries*. In FedCSIS Communication Papers (pp. 197-204).

Bray, F., Parkin, D. M. (2009). *Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness*. European journal of cancer, 45(5), 747-755.

Cai, L., Zhu, Y. (2015). *The challenges of data quality and data quality assessment in the big data era*. Data science journal, 14.

Caro, A., Calero, C., Piattini, M. (2007). *A Portal Data Quality Model For Users And Developers*. In ICIQ (pp. 462-476).

Chen, D., Asaolu, B., Qin, C. (2016). *Big Data Analytics In The Public Sector: a Case Study of Neet Analysis For The London Boroughs*. IADIS International Journal on Computer Science & Information Systems, 11(2).

Chungoora, N., Young, R. I., Gunendran, G., Palmer, C., Usman, Z., Anjum, N. A., ... Case, K. (2013). *A model-driven ontology approach for manufacturing system interoperability and knowledge sharing*. Computers in Industry, 64(4), 392-401.

Colpaert, P., Joye, S., Mechant, P., Mannens, E., Van de Walle, R. (2013). *The 5 stars of open data portals*. In Proceedings of the 7th International Conference on Methodologies, Technologies and Tools Enabling E-Government (MeTTeG13), University of Vigo, Spain (pp. 61-67).

Coutinho, C., Cretan, A., & Jardim-Goncalves, R. (2012). *Negotiations framework for monitoring the sustainability of interoperability solutions*. In International IFIP Working Conference on Enterprise Interoperability (pp. 172-184). Springer, Berlin, Heidelberg.

Dahbi, K. Y., Lamharhar, H., Chiadmi, D. (2018). *Toward an Evaluation Model for Open Government Data Portals*. In International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning (pp. 502-511). Springer, Cham.

Färber, M., Bartscherer, F., Menne, C., Rettinger, A. (2018). *Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago*. Semantic Web, 9(1), 77-129.

Ferney, M. M. J., Estefan, L. B. N., Alexander, V. V. J. (2017). *Assessing data quality in open data: A case study*. In 2017 Congreso Internacional de Innovacion y Tendencias en Ingenieria (CONIITI) (pp. 1-5). IEEE.

Fisher, C. W., Kingma, B. R. (2001). *Criticality of data quality as exemplified in two disasters*. Information & Management, 39(2), 109-116.

Friedman, T., Judah, S. (2013). *The state of data quality: Current practices and evolving trends*. Stamford: Gartner.

Friedman, T., Smith, M. (2011). *Measuring the business value of data quality*. Gartner, Stamford, 464.

Guha-Sapir, D., Below, R. (2002). *The quality and accuracy of disaster data: a comparative analyses of three global datasets*. World Bank, Disaster Management Facility, ProVention Consortium.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., Khan, S. U. (2015). *The rise of "Big data" on cloud computing: Review and open research issues*. Information systems, 47, 98-115.

Haubold, T., Beier, G., Golubski, W., Herbig, N. (2010). *The GeneSEZ approach to model-driven software development*. In Advanced Techniques in Computing Sciences and Software Engineering (pp. 395-400). Springer, Dordrecht.

Jetzek, T. (2017). *Innovation in the open data ecosystem: Exploring the role of real options thinking and multi-sided platforms for sustainable value generation through open data*. In Analytics, Innovation, and Excellence-Driven Enterprise Sustainability (pp. 137-168). Palgrave Macmillan, New York.

Kerr, K., Norris, T. (2007). *The development of a health data quality programme*. In Information quality management: Theory and applications (pp. 94-118). IGI Global.

Kessler, C. W., Schamai, W., Fritzson, P. (2010, February). *Platform-independent modeling of explicitly parallel programs*. In 23th International Conference on Architecture of Computing Systems 2010 (pp. 1-11). VDE.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Klein, R. H., Klein, D. B., Luciano, E. M. (2018). *Open Government Data: Concepts, Approaches and Dimensions over Time*. Revista Economia & Gestão, 18(49), 4-24.

Kleppe, A. (2008). *Software language engineering: creating domain-specific languages using metamodels*. Pearson Education.

Kleppe, A. G., Warmer, J., Warmer, J. B., Bast, W. (2003). *MDA explained: the model driven architecture: practice and promise*. Addison-Wesley Professional.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A. (2014). *Test-driven evaluation of linked data quality*. In Proceedings of the 23rd international conference on World Wide Web (pp. 747-758). ACM.

Kučera, J., Chlapek, D., Nečaský, M. (2013). *Open government data catalogs: Current approaches and quality perspective*. In International conference on electronic government and the information systems perspective (pp. 152-166). Springer, Berlin, Heidelberg.

Kuk, G., Davies, T. (2011). *The roles of agency and artifacts in assembling open data complementarities*.

Lano, K. (2005). *Advanced systems design with Java, UML and MDA*. Elsevier.

Larsen, I. K., Småstuen, M., Johannesen, T. B., Langmark, F., Parkin, D. M., Bray, F., Møller, B. (2009). *Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness*. European journal of cancer, 45(7), 1218-1231.

Lee, Y. W., Pipino, L. L., Funk, J. D., Wang, R. Y. (2009). *Journey to data quality*. The MIT Press.

Mellor, S. J., Scott, K., Uhl, A., Weise, D. (2004). *MDA distilled: principles of model-driven architecture*. Addison-Wesley Professional.

Miller, J., Mukerji, J. (2003). *MDA Guide Version 1.0. 1*. Object Management Group, 234, 51.

Moore, S. (2017). *How to Create a Business Case for Data Quality Improvement*. Retrieved January, 27, 2018.

Neumaier, S. (2015). *Open Data Quality: Assessment and Evolution of (Meta-) Data Quality in the Open Data Landscape*. Technische Universität Wien.

Nikiforova, A. (2020a). *Comparative analysis of national open data portals or whether your portal is ready to bring benefits from open data*. In IADIS International Conference on ICT, Society and Human Beings 2020, Part of the IADIS Multi Conference on Computer Science and Information Systems, MCCSIS 2020, July 21 - 23, 2020. IADIS.

Nikiforova, A. (2020b). *Assessment of Latvia's Open Data Portal or how Close are We to Gaining Benefits from Open Data*, In 14th International Conference on Interfaces and Human Computer Interaction, Part of the IADIS Multi Conference on Computer Science and Information Systems, MCCSIS 2020, July 23 - 25, 2020. IADIS.

Nikiforova, A. (2020c). *Datu kvalitātes definēšana un novērtēšana. Definition and evaluation of data quality* (Doctoral Thesis).

Nikiforova, A. (2019a). *Analysis of open health data quality using data object-driven approach to data quality evaluation: insights from a Latvian context*. In IADIS International Conference e-Health 2019, Part of the IADIS Multi Conference on Computer Science and Information Systems, MCCSIS 2019, July 16 - 19, 2019 (pp. 119-126). IADIS.

Nikiforova, A. (2018a). O*pen Data Quality Evaluation: A Comparative Analysis of Open Data in Latvia*. Baltic Journal of Modern Computing, 6(4), 363-386.

Nikiforova, A. (2018b). *Open Data Quality*. In Doctoral Consortium/Forum@ DB&IS (pp. 151-160)

Nikiforova, A., Bicevska, Z. (2018). Application of LEAN Principles to Improve Business Processes: a Case Study in Latvian IT Company. Baltic Journal of Modern Computing, 6(3), 247-270.

Nikiforova, A., Bicevskis, J. (2020). *Towards a Business Process Model-based Testing of Information Systems Functionality*. In Proceedings of the 22nd International Conference on Enterprise Information Systems – Vol. 2: *ICEIS*, ISBN 978-989-758-423-7, p. 322-329. DOI: 10.5220/0009459703220329

Nikiforova, A., Bicevskis, J., Bicevska, Z., Oditis, I. (2020). User-Oriented Approach to Data Quality Evaluation. Journal of Universal Computer Science, 26(1), 107-126.

Nikiforova, A., Bicevskis, J. (2019). *An Extended Data Object-driven Approach to Data Quality Evaluation: Contextual Data Quality Analysis*. Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019), 274-281. DOI: 10.5220/0007838602740281

Olson, J. E. (2003). *Data quality: the accuracy dimension*. Elsevier.

Ostadzadeh, S. S., Aliee, F. S., Ostadzadeh, S. A. (2008). *An MDA-based generic framework to address various aspects of enterprise architecture*. In Advances in Computer and Information Sciences and Engineering (pp. 455-460). Springer, Dordrecht.

Parkin, D. M., Bray, F. (2009). *Evaluation of data quality in the cancer registry: principles and methods Part II*. Completeness. European journal of cancer, 45(5), 756-764.

Pauker, F., Frühwirth, T., Kittl, B., Kastner, W. (2016). *A systematic approach to OPC UA information model design*. Procedia CIRP, 57, 321-326.

Paulheim, H., Bizer, C. (2014). *Improving the quality of linked data using statistical distributions*. International Journal on Semantic Web and Information Systems (IJSWIS), 10(2), 63-86.

Perez-Castillo, R., Carretero, A. G., Caballero, I., Rodriguez, M., Piattini, M., Mate, A., ... Lee, D. (2018a). DAQUA-MASS: An ISO 8000-61 based data quality management methodology for sensor data. Sensors, 18(9), 3105.

Perez-Castillo, R., Carretero, A. G., Rodriguez, M., Caballero, I., Piattini, M., Mate, A., ... Lee, D. (2018b). Data quality best practices in IoT environments. In *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)* (pp. 272-275). IEEE.

Petychakis, M., Vasileiou, O., Georgis, C., Mouzakitis, S., Psarras, J. (2014). *A state-of-the-art analysis of the current public data landscape from a functional, semantic and technical perspective*. Journal of theoretical and applied electronic commerce research, 9(2), 34-47.

Price, R., Shanks, G. (2016). *A semiotic information quality framework: development and comparative analysis*. In Enacting Research Methods in Information Systems (pp. 219-250). Palgrave Macmillan, Cham.

Redman, T. C. (2001). *Data quality: the field guide*. Digital press.

Ruijer, E., Meijer, A. (2019). *Open Government Data as an Innovation Process: Lessons from a Living Lab Experiment*. Public Performance & Management Review, 1-23.

Ruiz, M. (2018). TraceME: *A Traceability-Based Method for Conceptual Model Evolution*. Springer International Publishing.

Sáez Martín, A., Rosario, A. H. D., Pérez, M. D. C. C. (2016). *An international analysis of the quality of open government data portals*. Social Science Computer Review, 34(3), 298-311.

Sasse, T., Smith, A., Broad, E., Kennison, J., Wells, P., Atz, U. (2017) "*Recomendations for Open Data Portals: from Setup to sustainability*", Disponível na WWW: https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

Scannapieco, M., Missier, P., Batini, C. (2005). *Data quality at a glance*. Datenbank-Spektrum, 14(January), 6-14.

Scannapieco, M., Catarci, T. (2002a). *Data quality under a computer science perspective*. Archivi & Computer, 2, 1-15.

Scannapieco, M., Pernici, B., Pierce, E. M. (2002b). *IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework*. In ICIQ (pp. 279-291).

Selic, B. (2009). *The theory and practice of modeling language design for model-based software engineering—a personal perspective*. In International Summer School on Generative and Transformational Techniques in Software Engineering (pp. 290-321). Springer, Berlin, Heidelberg.

Shi, X., Han, W., Huang, Y., Li, Y. (2005). *Service-oriented business solution development driven by process model*. In The Fifth International Conference on Computer and Information Technology (CIT'05) (pp. 1086-1092). IEEE.

Sigurdardottir, L. G., Jonasson, J. G., Stefansdottir, S., Jonsdottir, A., Olafsdottir, G. H., Olafsdottir, E. J., Tryggvadottir, L. (2012). *Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness*. Acta oncologica, 51(7), 880-889.

Soley, R. (2000). *Model driven architecture*. OMG white paper, 308(308), 5.

Sprogis, A., Barzdins, J. (2013). *Specification, Configuration and Implementation of DSL Tool*. In Databases and Information Systems VII: Selected Papers from the Tenth International Baltic Conference, DB & IS 2012 (Vol. 249, p. 330). IOS Press.

Schmidt, M., Schmidt, S. A. J., Sandegaard, J. L., Ehrenstein, V., Pedersen, L., Sørensen, H. T. (2015). *The Danish National Patient Registry: a review of content, data quality, and research potential*. Clinical epidemiology, 7, 449.

Tayi, G. K., Ballou, D. P. (1998). *Examining data quality*. Communications of the ACM, 41(2), 54-57.

Tinholt, D. (2013). *The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data*. Capgemini Consulting.

Tomic, K., Sandin, F., Wigertz, A., Robinson, D., Lambe, M., & Stattin, P. (2015). *Evaluation of data quality in the National Prostate Cancer Register of Sweden*. European journal of cancer, 51(1), 101-111.

Umbrich, J., Neumaier, S., Polleres, A. (2015, March). *Towards assessing the quality evolution of open data portals*. In Proceedings of ODQ2015: Open Data Quality: from Theory to Practice Workshop, Munich, Germany.

Van den Berghe, S., Van Gaeveren, K. (2017). *Data quality assessment and improvement: a Vrije Universiteit Brussel case study*. Procedia Computer Science, 106, 32-38.

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., Morando, F. (2016). *Open data quality measurement framework: Definition and application to Open Government Data*. Government Information Quarterly, 33(2), 325-337.

Wang, R. Y., Strong, D. M. (1996). *Beyond accuracy: What data quality means to data consumers*. Journal of management information systems, 12(4), 5-33.

Weiskopf, N. G., Weng, C. (2013). *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research*. Journal of the American Medical Informatics Association, 20(1), 144-151.

Yi, M. (2019). *Exploring the quality of government open data: Comparison study of the UK, the USA and Korea*. The Electronic Library, 37(1), 35-48.

Zhao, W., Bryant, B. R., Raje, R. R., Auguston, M., Gray, J. G., Burt, C. C., Olson, A. M. (2003). *A generative and model driven framework for automated software product generation*. Alabama Univ in Birmingham Dept Of Computer and Information Sciences.

WEB (a). *Global Open Data Index*. https://index.okfn.org/

WEB (b). *The "All In" Costs of Poor Data Quality. It goes beyond dollars and cents*, ComputerWorld, 2015, https://www.computerworld.com/article/2949323/the-all-in-costs-of-poor-data-quality.html

WEB (c). *The Ultimate Guide to Modern Data Quality Management (DQM) For An Effective Data Quality Control Driven by The Right Metrics*, The Data Pine Blog, 2018, https://www.datapine.com/blog/data-quality-management-and-metrics/

WEB (d). *Inside Big Data. The Hidden Costs of Bad Data, 2017*, https://insidebigdata.com/2017/05/05/hidden-costs-bad-data/

WEB (e). *Ten Principles for Opening Up Government Information*, https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

Received June 5, 2020, accepted June 8, 2020 as a reviewed paper