An Extended Data Object-driven Approach to Data Quality Evaluation: Contextual Data Quality Analysis

Anastasija Nikiforova[a] and Janis Bicevskis[b]
Faculty of Computing, University of Latvia, 19 Raina Blvd., Riga, Latvia

Keywords: Data Quality, Contextual Control, Data Object, Executable Models, Domain-specific Modelling Language.

Abstract: This research is an extension of a data object-driven approach to data quality evaluation allowing to analyse data object quality in scope of multiple data objects. Previously presented approach was used to analyse one particular data object, mainly focusing on syntactic analysis. It means that the primary data object quality can be analysed against secondary data objects of unlimited number. This opportunity allows making more comprehensive, in-depth contextual data object analysis. The given analysis was applied to open data sets, making comparison between previously obtained results and results of application of the extended approach, underlying importance and benefits of the given extension.

1 INTRODUCTION

Data quality has been a topical issue for many decades, and it has a growing importance nowadays for several reasons: (a) in the 21st century data is everywhere, according to (Economist, 2017) “the world’s most valuable resource is no longer oil, but data”; (b) open data popularity increases extremely fast – a large number of open data portals and data sets have become available all over the world.

According to (OpenDataSoft, 2019), there are more than 2600 open data portals around the world already and these numbers continuously increase every day.

The increasing amount of data triggered the data quality issue. The most significant aspect of data quality is its impact on decision-making, thus the use of low-quality data may cause misleading decisions. Usually the open data are processed without implying any quality checks. However, statistics on losses inflicted by data quality problems are frustrating. According to (Gartner, 2013) and (Gartner, 2018), the use of the poor data quality results in $15 million to $3.1 trillion dollars of losses on the US Economy per year.

In previous research (Bicevskis et al., 2018b) the initial version of data object-driven approach to data quality evaluation was proposed. It consists of 3 components: data object, quality specification, quality measuring process. Data quality evaluation for the specific purpose requires description of the requirements for data values that should be executable, since the stored data are “scanned” and checked for compliance with the requirements. The proposed solution requires the development of a two-level data quality assurance procedure:

(1) syntactic control – values of data objects are checked locally within one record (compliance of input data with the syntax);

(2) semantic/ contextual control – checking whether the newly entered data is compatible with the previously entered data that is stored in database. The semantic control should be repeated every time when entering or editing the new values of data objects’ attributes.

Syntactic control was already described in (Bicevskis et al., 2018a, 2018b), (Nikiforova, 2018a). However, as data quality analysis was made in scope of one data object and the focus lied on analysis of syntax, it wasn’t complete and deep enough. In this paper the presented approach was generalized and used for semantic data quality analysis. Unlike the syntactic quality analysis, the semantic quality allows defining the context of other/external/secondary data objects used to analyse the primary object.

[a] https://orcid.org/0000-0002-0532-3488
[b] https://orcid.org/0000-0001-5298-9859

Nikiforova, A. and Bicevskis, J.
An Extended Data Object-driven Approach to Data Quality Evaluation: Contextual Data Quality Analysis.
DOI: 10.5220/0007838602740281
In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 274-281
ISBN: 978-989-758-372-8

The paper deals with the following issues: a short overview of the related researches (Section 2), a rationale for the proposed solution (Section 3), a description of the proposed solution (Section 4), an application of the proposed solution to open data sets and comparative evaluation with previous results (Section 5), conclusions and future work (Section 6).

2 RELATED RESEARCH

Previous papers of the authors (Nikiforova, 2018a) and (Bicevskis et al., 2018a) provide an overview of 20 different data quality solutions that were proposed in last decade, emphasizing their pros and cons, so this research will not explain and discuss them in detail. In addition, in (Batini et al., 2016) and (Batini et al., 2009) a comprehensive overview of methodologies for data quality assessment from different perspectives is provided.

One of the most traditional concepts of data quality analysis that is used by the most part of the existing researches is data quality dimension. The most popular and commonly used among existing sets of data quality dimensions are (a) Wang's and Strong's 15 data quality dimensions representing four categories: intrinsic, contextual, representational, accessibility; (Wang and Strong, 1996), (b) Redman's 3 categories of data quality dimensions such as conceptual schema, data values and data format (Redman, 1997), (Batini et al., 2016), and (c) one of the most modern - the 6 data quality dimensions defined by Data Management Association International UK Working Group (DAMA, 2019).

However, there is no consensus on data quality dimensions and their usability. According to (Batini et al., 2016) "dimensions are not defined in a measurable and formal way", other authors are of the same opinion (DAMA, 2019), (Huang et al., 1999), (Eppler, 2006). According to (DAMA, 2019), "even amongst data quality professionals the key data quality dimensions are not universally agreed. This state of affairs has led to much confusion within the data quality community and is even more bewildering for those who are new to the discipline and more importantly to business stakeholders".

Despite the fact that many data quality dimensions and their groupings have already been proposed by well-known researchers, the most of nowadays data quality researches are focused on finding of new, "more comprehensive" dimensions. One of such examples is the PoDQA project (Caro et al., 2007) - a quite promising data quality model for Web portals that uses 30 data quality attributes.

Moreover, many solutions propose using data quality dimensions of the same name but with different semantics and vice versa – different names for the same dimensions (Batini et al., 2016).

Existing researches intend that the data quality requirements (defined by a user or already existing ones) are assigned to appropriate dimensions to be further applied for data sets. But it leads to the necessity to involve data quality experts at every stage of data quality analysis process as very deep and specific knowledges of data quality and data quality dimensions are required. However, the presented data object-driven approach eliminates this very specific, time-consuming and in most cases confusing step, using more comprehensive concept "data quality requirement". Quality requirements are defined for the specific data object without their assigning to the specific data quality dimension and group, as the data quality is a relative concept that depends on the use-case. For instance, if a user exploits the data set "Company Register" identifying company by its name, registration number and incorporation date, he might not be interested in information about representative persons of the company. However, another user may use the same data set to contact the company through its representative person, in this case the data of the company's representants could be of big importance but incorporation date could be completely useless. It means the same data set can be of high quality for one user and completely useless (of low quality) for another. The proposed approach recommends defining the data object which quality will be analysed first. The data quality requirements are defined depending on the specific use-case for the previously defined data object.

To sum up, there are no well-known data quality solutions which would allow non-data quality staff to analyse "foreign"/ "external" data sets without any information about how data was initially collected and processed. The proposed solution is expected to provide possibility to analyse data sets according to the needs of users, i.e., to specific use-cases.

3 RATIONALE FOR THE RESEARCH

With the popularity of open data, the need for quality testing is also increasing. In accordance with the nature of open data, its data quality analysis should be simple enough for users without advanced IT and data quality knowledges. In order to ensure users with such possibility, very intuitive data object-driven approach to the data quality evaluation was previously presented in (Bicevskis et al., 2018a). It was designed to analyse parameters of one data object, mainly focusing on syntactic analysis. According to the experience of its application to multiple data sets (Bicevskis et al., 2018b), (Nikiforova, 2018a, 2018b), the most common checks were focused on: (1) existence of values, (2) relevance to the specified data type, (3) relevance to the list of enumerable values, (4) validity of value (for example, credible date).

This paper considers the analysis of data object's quality in the context of multiple data objects out of scope. The necessity for data quality model in the context of multiple data objects is obvious as provided by the data quality analysis in (Bicevskis et al., 2018a). In previous research the data quality of Company Registers of four European countries was analysed. Two the most obvious use-cases were chosen: (a) identifying company by name, registration number and incorporation date; (b) contacting company via post using its address and postal code. Despite the very simple use-case the results of the data quality analysis were surprising – many data quality problems were detected. However, in many cases the single data object analysis indicated the mere existence of the data quality problem without detecting all the defective records.

For instance, analysis of Companies House of the United Kingdom (Companies House, 2018) showed significant number of data quality deficits in the fields storing information on country. More in-depth analysis was required involving the context of another data object containing information on country (conforming standards).

The proposed extension is also valuable analysing data sets of .json and .xml formats, as these formats can be treated as many interconnected tables, where number of tables depends on the document structure.

Contextual checks are often implemented in relational databases and IS (for instance, if the record already exists). They usually are checked after or together with syntactic checks. However, in other cases such as non-relational databases the quality checks are difficult to test and to manage. Users without having access to data storing, accruing and processing procedures, cannot check the data quality. To solve this problem, these checks (same as syntactic checks in the initial version of the presented approach) are separated from program code.

The creation of data quality models within the context of multiple data objects allows making more comprehensive in-depth data quality analysis.

4 DATA QUALITY MODEL

The proposed data object-oriented approach covers 3 of 4 data quality control phases proposed by Total Data Quality Management (TDQM, 2019): (1) data quality definition, (2) data quality measuring, and (3) data quality analysis. The 4th phase - data quality improvement – is left to users or data publishers as there exist many useful and user-friendly solutions for this purpose. As stated before (Bicevskis et al., 2018b), there are three main components forming the proposed data quality model: (a) data object that specifies the data which quality will be evaluated, (b) quality requirements that define conditions that must be met to admit data as qualitative, and (c) quality evaluation process that defines procedures should be performed to evaluate the data object's quality.

All three components are defined by using a graphical domain specific language (DSL). Three DSL families were developed as graphic languages (Nikiforova, 2018a) based on the possibilities of the modelling platform DIMOD (Bicevska et al., 2017). DIMOD is advised to be used instead of specific data objects definition language as it offers a possibility to define DSLs with different data objects structures.

Requirements for data quality syntaxis are logical conditions on data fields of the primary data objects. Requirements for data quality semantics are defined by logical conditions on data fields of the secondary data object. Therefore, both syntactical and semantical data quality can be analysed according to unified description principles.

One of the advantages of the proposed approach is that diagrams are easy to create, edit and read. Graphical diagrams (a) make the data quality analysis process easier and well-understandable to non-IT and non-data quality experts, (b) support interaction between users (between business- and technical-level units, different persons or even teams within an organization or many organizations if needed) as more than one person can be involved in the process.

According to the proposed approach, the language describing the quality evaluation process involves verification activities for individual data objects that can be defined in several ways: (a) informally as a natural language text, (b) using UML activity diagrams, or (c) in the own DSL.

More detailed definition of each component is provided in the next Section. Two examples explain main concepts, changes and improvements in comparison with previously defined (Bicevskis et al., 2018b): (a) data quality analysis in scope of one data set for The Register of Enterprises of the Republic of Latvia (2019); (b) data quality analysis in context of multiple data objects and data sets for Companies House (Company House, 2018).

4.1 Primary and Secondary Data Objects

The first and probably one of the most important concepts of the presented approach is a "data object". As previously stated in (Bicevskis et al., 2018b), "a data object is the set of values of the parameters that characterize a real-life object". It means that (a) data object stores only the data the particular user is interested in, (b) every analysis can have different attribute names and types. Figure 1 shows a single data object "Enterprise_LV" with its attributes that are necessary in specific use-cases (with just few additional attributes) to get an overview of Enterprise Register quality: Reg_number – registration number of enterprise, Name – name of enterprise, Type – type of enterprise etc. The description of data object at this stage is informal - no rules for attribute values syntax are given.

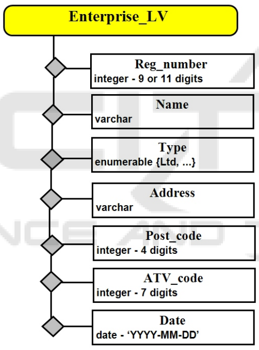


Figure 1: Single data object "Enterprise_LV".

Figure 2 demonstrates a data object in the context of multiple data objects. In this case, one object becomes the primary and other objects are secondary. The primary data object is the initial data object which quality is analysed. The secondary data object determines the context for analysis of the primary data object. The primary data object traditionally is one (data quality analysis for one data set), however the number of the secondary data objects is not limited, as it on the nature of the primary data object and its parameters. The number of secondary data objects depends on the primary data object and its parameters. A secondary data object is introduced if it is possible and necessary to analyse the data quality of the primary data object against another data object. Moreover, (a) more than one secondary data object can be related to any parameter of the primary data object, (b) every secondary data object can be related to multiple parameters of the primary data object.

In the example, the quality of the primary data object "Company_UK" is analysed against the secondary data object "Country" that store codes for the representation of countries' names and their subdivisions (according to ISO 3166, including such popular standards as ISO2, ISO, Short name, Official name, UNDP). In the Figure 2, two parameters of the primary data object "Company_UK" ("Companies House of United Kingdom") storing country names ("RegAddress Country" and "CountryOfOrigin") are connected with all parameters of the secondary data object "Country".

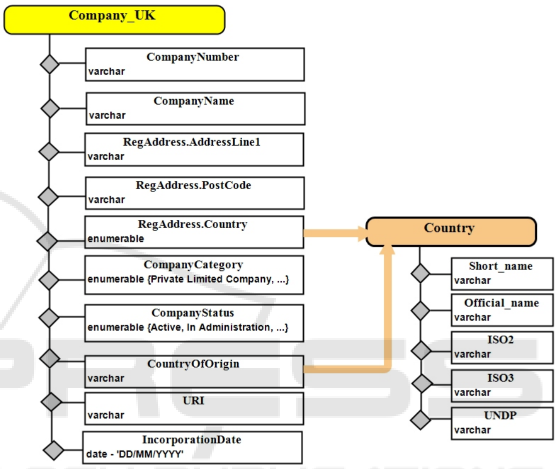


Figure 2: Data object in context of multiple data objects.

Every secondary data object has its own colour that is also assigned to arrows connecting the primary and the secondary data objects. As the description of data object at this stage is informal, an arrow also doesn't represent any semantics – it just depicts an interconnection of primary and secondary data objects.

It is also possible to add supplementary data objects to connect the primary and the secondary data objects if needed. However, this paper discusses the most common situation when data objects can be directly connected by one or more parameter's values.

4.2 Data Object Quality Specification

The aim of data quality specification is to allow a user to define conditions that must be met to admit the data object (that was defined on the previous stage) as of high quality. There are two possible ways to define quality conditions: (a) informal descriptions of conditions, for example, in a natural language, or (b) formalized descriptions that are implementation independent. According to the approach (Bicevskis et al., 2018b), “data quality specification of a data object is defined by logical expressions, where names of data object’s attributes/ fields serve as operands in logical expressions”.

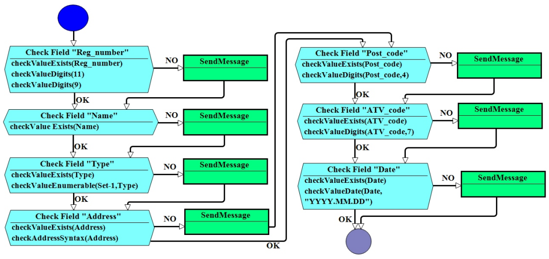


Figure 3: Data quality specification.

In case of multiple data objects, additional two steps at this stage should be taken: (1) supplying data quality specification diagram (Figure 3) with secondary data object(-s), and (2) specifying how the primary and the secondary data objects are linked/ connected (Figure 4).

Quality conditions are defined only for the primary data object (“Company_UK” in Figure 4). Data quality of the secondary data object (“Country”) isn’t analysed. Only quality of the primary data objects is analysed in case of contextual data quality analysis, as the secondary data objects are auxiliary, they just supplement analysis. It is more convenient to ensure data quality of the secondary data object before it is involved in data quality analysis of another data object. If the secondary data object quality isn’t checked before, resulting records will contain records, where part of them will point out quality problems of the second data object but detection of the quality defects of the primary data object will be more complicated and resource consuming. Moreover, some quality problems won’t be found at all, if the quality of the secondary data object is not high enough. Since the secondary data object’s quality isn’t analysed, no quality conditions are defined for the data object “Country” (Figure 4). The initial structure of the data quality specification diagram is just supplemented with the secondary data object as shown in the Figure 4.

When the secondary data object is specified, links between primary and secondary data objects must be depicted. At this stage, informal rules could be added to describe the connections. They are depicted by adding arrows linking both (a) parameter of the primary data object which quality is checked against the secondary data object, and (b) parameters of the secondary data object involved in this check. The number of attributes of the secondary data object involved in the specific quality check determines the number of dash lines connecting the arrow and the secondary data object.

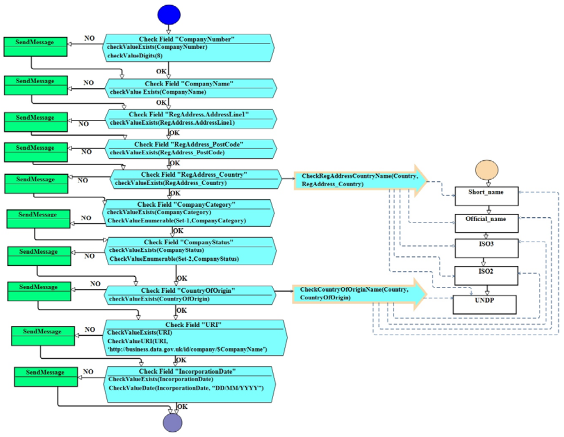

SendMessage
NO
Check Field "CompanyNumber"
checkValueExists(CompanyNumber)
checkValueDigits(8)
OK
SendMessage
NO
Check Field "CompanyName"
checkValue Exists(CompanyName)
OK
SendMessage
NO
Check Field "RegAddress.AddressLine1"
checkValueExists(RegAddress.AddressLine1)
OK
SendMessage
NO
Check Field "RegAddress_PostCode"
checkValueExists(RegAddress_PostCode)
OK
SendMessage
NO
Check Field "RegAddress_Country"
checkValueExists(RegAddress_Country)
CheckRegAddressCountryName(Country,
RegAddress_Country)
OK
SendMessage
NO
Check Field "CompanyCategory"
checkValueExists(CompanyCategory)
CheckValueEnumerable(Set-1,CompanyCategory)
OK
SendMessage
NO
Check Field "CompanyStatus"
checkValueExists(CompanyStatus)
CheckValueEnumerable(Set-2,CompanyStatus)
OK
SendMessage
NO
Check Field "CountryOfOrigin"
checkValueExists(CountryOfOrigin)
CheckCountryOfOriginName(Country,
CountryOfOrigin)
OK
SendMessage
NO
Check Field "URI"
CheckValueExists(URI)
CheckValueURI(URI,
'http://business.data.gov.uk/id/company/$CompanyName')
OK
SendMessage
NO
Check Field "IncorporationDate"
CheckValueExists(IncorporationDate)
CheckValueDate(IncorporationDate, "DD/MM/YYYY")
OK
Short_name
Official_name
ISO3
ISO2
UNDP

Figure 4: Data quality specification in context of multiple data objects.

In the given example (Figure 4), the quality of two parameters (“RegAddress_Country” and “CountryOfOrigin”) is analysed against all parameters of the secondary data object. It means all parameters of the secondary data object “Country” are connected by the arrow containing conditions and the dashed lines. The colour of the secondary data object input element corresponds to the colour assigned to this data object at the 1st stage. Informal conditions added to arrows are formulated in the same way as the conditions for each specific parameter. The arrow’s colour corresponds to the box containing requirements while the border’s colour corresponds to the secondary data object’s colour.

Because of data quality’s relative nature, data quality requirements depend on users’ needs, i.e. on the defined use-cases. In the given example, several data quality conditions are possible: (1) all country names available in the primary data object must conform to at least one standard; (2) all country names available in the primary data object must conform to the one particular standard. In the second case, following options are also possible: (a) conformity to one specific standard defined by a specific user, (b) conformity to one of widely-used standards. In later case, data object’s pre-processing is required to determine the standard which the most records correspond to. The determined standard is used in quality requirements.

The type of requirement depends on user’s definition of the “data quality” concept. The first option can be used to check whether the stored values are valid (representing an object of the real world), whilst the second option is to check whether all data conforms to the same standards (sometimes called “homogeneity”).

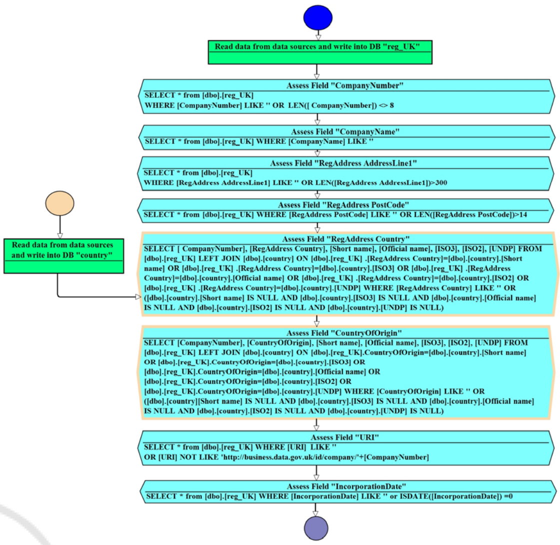
4.3 Quality Evaluation Process

Quality evaluation process consists of two steps: (1) the first step describes activities to be taken to select data object values from data sources (data object class instances are selected from the data sources and written into a collection), (2) one or more steps are to be taken to evaluate quality of primary data objects according to quality specifications. As stated in (Bicevskis et al., 2018b), all instances are processed cyclically by examining the fulfilment of a quality specification for each individual instance.

Analysing data object in context of multiple objects, the number of data objects involved in the analysis determines the number of data sources where data object values must be selected from. Data object values of the secondary data objects are gathered from the data source if the parameter indicating the secondary data object’s value in scope of defined quality requirement/condition is true (Figure 5). The following amendments should be done: (a) depiction of data read/ write operations from data source into database, (b) connection via appropriate parameters, (c) appropriate quality requirements that are depicted in the primary data object’s box together with other conditions.

At this stage diagram should contain separate field checks for the primary data object where each individual operation evaluates the data quality of the field by using a SQL statement, programming code routines or similar executable artefacts (depending on person’s knowledges and experience).

In this example, SQL statements are used as the most intuitive option. Initially only the SQL (a) SELECT statement that specifies the target data object, and (b) WHERE clause that specifies the quality specification, were needed. However, the secondary data objects involved in this process require additional clause that would link/ combine the primary and the secondary objects based on one or more related parameters between them. This task in SQL is solved by JOIN clause. Type of the JOIN used to combine data objects depends on the implementation (the sequence in which they appear in the statement): LEFT is used if the secondary data object appears the first (Figure 5), and RIGHT if vice versa. It means that only three basic SQL statement’s components form a data quality requirement’s/ condition’s structure for every single analysed parameter. However, from case to case more complex constructions and SQL operators are needed for translating informally defined data quality requirements into executable statements.


Read data from data sources and write into DB "reg_UK"
Read data from data sources and write into DB "country"
Assess Field "CompanyNumber"
Assess Field "CompanyName"
Assess Field "RegAddress AddressLine1"
Assess Field "RegAddress PostCode"
Assess Field "RegAddress Country"
Assess Field "CountryOfOrigin"
Assess Field "URI"
Assess Field "IncorporationDate"

```
SELECT * from [dbo].[reg_UK]
WHERE [CompanyNumber] LIKE '' OR LEN([ CompanyNumber]) <> 8
```

```
SELECT * from [dbo].[reg_UK] WHERE [CompanyName] LIKE ''
```

```
SELECT * from [dbo].[reg_UK]
WHERE [RegAddress AddressLine1] LIKE '' OR LEN([RegAddress AddressLine1])>300
```

```
SELECT * from [dbo].[reg_UK] WHERE [RegAddress PostCode] LIKE '' OR LEN([RegAddress PostCode])>14
```

```
SELECT [ CompanyNumber], [RegAddress Country], [Short name], [Official name], [ISO3], [ISO2], [UNDP] FROM
[dbo].[reg_UK] LEFT JOIN [dbo].[country] ON [dbo].[reg_UK] .[RegAddress Country]=[dbo].[country].[Short
name] OR [dbo].[reg_UK] .[RegAddress Country]=[dbo].[country].[ISO3] OR [dbo].[reg_UK] .[RegAddress
Country]=[dbo].[country].[Official name] OR [dbo].[reg_UK] .[RegAddress Country]=[dbo].[country].[ISO2] OR
[dbo].[reg_UK] .[RegAddress Country]=[dbo].[country].[UNDP] WHERE [RegAddress Country] LIKE '' OR
([dbo].[country].[Short name] IS NULL AND [dbo].[country].[ISO3] IS NULL AND [dbo].[country].[Official name]
IS NULL AND [dbo].[country].[ISO2] IS NULL AND [dbo].[country].[UNDP] IS NULL)
```

```
SELECT [CompanyNumber], [CountryOfOrigin], [Short name], [Official name], [ISO3], [ISO2], [UNDP] FROM
[dbo].[reg_UK] LEFT JOIN [dbo].[country] ON [dbo].[reg_UK].CountryOfOrigin=[dbo].[country].[Short name]
OR [dbo].[reg_UK].CountryOfOrigin=[dbo].[country].[ISO3] OR
[dbo].[reg_UK].CountryOfOrigin=[dbo].[country].[Official name] OR
[dbo].[reg_UK].CountryOfOrigin=[dbo].[country].[ISO2] OR
[dbo].[reg_UK].CountryOfOrigin=[dbo].[country].[UNDP] WHERE [CountryOfOrigin] LIKE '' OR
([dbo].[country][Short name] IS NULL AND [dbo].[country].[ISO3] IS NULL AND [dbo].[country].[Official name]
IS NULL AND [dbo].[country].[ISO2] IS NULL AND [dbo].[country].[UNDP] IS NULL)
```

```
SELECT * from [dbo].[reg_UK] WHERE [URI]  LIKE ''
OR [URI] NOT LIKE 'http://business.data.gov.uk/id/company/'+[CompanyNumber]
```

```
SELECT * from [dbo].[reg_UK] WHERE [IncorporationDate] LIKE '' or ISDATE([IncorporationDate]) =0
```

Figure 5: Data quality evaluation process (in context of multiple data objects).

Figure 5 shows how the parameters requiring contextual data quality checks against the secondary data objects are highlighted using border colour corresponding to the colour was assigned to the appropriate secondary data object. Each secondary data object is represented by using the same colour in all quality evaluation stages (in all three diagrams).

5 ANALYSIS RESULTS

The extended approach was examined and approbated through applying it to 14 data sets. In this paper we explain step-by-step analysis of the previously mentioned data set “Companies House of the United Kingdom” in order to compare result of this analysis with previous results (Bicevskis et al., 2018a). Two parameters containing name of the country “CountryOfOrigin” and “RegAddress Country” are validated against the secondary data object “Country” storing the name of the country according to the international standards.

These parameters were chosen based on the results of the approbated initial version of the proposed approach (Bicevskis et al., 2018a) where the following data quality problems were detected: (1) various names indicating the same country (UK and United Kingdom; USA, United States and United

States of America etc.); (2) names of dissolved countries such as Czechoslovakia, Yugoslavia, USSR; (3) values indicating administrative division or region (Wales, Scotland, England & Wales, England); (4) 4 values are not countries at all: “SW7” - South Kensington and part of Knightsbridge postal code, “EAST SUSSEX” - a county in South East England, “BWI” - Baltimore/ Washington International Thurgood Marshall Airport code, “DE 19901” - postal code of Dover.

In total, 128 different values, that possibly can contain data quality problems, were indicated. The detection of these 128 values was quite sophisticated requiring multiple additional queries (grouping and sorting) in order to detect data anomalies indicating potential data quality problems. Therefore, the further analysis of these particular results would be very resource consuming. Moreover, there was lack of confidence that all possible data quality problems were detected and can be detected in this list at all.

These results indicated the necessity to make an additional analysis, which would (a) facilitate detection of these data quality problems, and (b) would provide only defective records which quality should be improved. It was achieved by presented extension.

Approbation of the proposed extended version of the approach shows that 33 of 109 (30%) country name records of the primary data object does not conform to any standard. This data quality problem was detected in 1 045 records (0.14%). 40 of 114 (35%) country name records of “RegAddress Country” parameter have data quality problems, i.e., don’t conform to any standard. This problem was detected in 206 994 records (27.44%).

Making an additional check whether the primary data object’s values conform to number of standards, it was concluded that: (a) all values of “CountryOfOrigin” conform to one standard, i.e., the short name, (b) 73 of 74 values of “RegAddress Country” conform to the same standard, however 1 value – “USA” corresponds to ISO3/UNDP.

Although this problem was observed in 208 039 (27.6%) records, it could be solved by making just 33 corrections in the first analysed parameter and 40 corrections in the second parameter. The number of necessary corrections was rather small as invalid names of both parameters overlap. Therefore, 48 values of country name should be corrected to reduce the detected problem. Correction of these 48 values would significantly improve the overall data object’s quality, reducing data quality problem in 208 039 records. The correction of 48 values does not require much of the resources (in comparison with 208 039).

In comparison, the extended version of the proposed approach yielded results that detected the records with the certain data quality problem. Comparing both lists it was concluded that the initial analysis detected 13 values with data quality problem instead of 48. 115 values didn’t have data quality problems. In this specific case, the proposed extended approach was able to improve the results of analysis by 72.9%. Therefore, the time needed to analyse the results of the initial approach now could be spent on the data quality improvement because the number of corrections is significantly lower than the number of resulted records where only 10.2% of values had data quality problems.

This example demonstrated the potential of the extended data object-driven approach. The number of possible controls where the proposed extended approach can yield valuable results is very high. The results of the contextual control can be used for data quality improvement.

Moreover, in case of open data, user’s participation in its quality analysis brings benefits not only the users themselves, but also data holders if users share their feedback. It can significantly improve not only users experience in its use but also the overall quality of open data.

6 CONCLUSIONS

The paper is a continuation and refinement of the authors’ previous researches in data quality area (Bicevskis, 2018b), (Nikiforova, 2018a). The paper focuses on the proposed data object-driven approach to data quality evaluation improving it with the possibility to analyse real data object’s quality within the context of multiple data objects. The extended approach yielded significantly improved results of the data quality analysis.

Moreover, although the extended approach was complemented with more in-depth analysis, the diagram’s structure remained easy to read, create, understand and edit even by non-IT and non-Data Quality professionals. It means that even deeper data quality analysis within the context of multiple data objects requires minimal involvement of IT-experts.

The presented extended approach is very intuitive. It significantly facilitates use of the data quality analysis by a regular user. The proposed structure eliminated the necessity of additional in-depth quality analysis, as well as writing complex queries and individual analysis of the results. The initial version of approach indicated records potentially containing data quality problems, however, this was very resources consuming process. The proposed extension of the approach detects only the records with the certain data quality problem by analysing the primary data object against other data objects that probably was collected by other data publishers. It allows to make more comprehensive and deep data quality analysis thus improving decision-making.

Although this extended approach was exemplified by examining open data sets, it is not specifically tailored to open data. It can also be applied to assess the quality of any type of structured or semi-structured data which is clear benefit of this extended approach. The authors expect that the proposed extended approach will provide possibility to analyse "foreign"/ "external" data sets without any information about initial data collection and processing. This will enable users to analyse any available data set according to their needs in the specific use-case.

Further research will be focused on (a) applying and evaluating the extended approach in the cases of complex data object's structure, including supplementing data objects in cases when direct connection between the primary and the secondary data objects is not possible, (b) detecting possible limitations of the proposed extended approach, (c) ensuring possibility to evaluate data sets' evolution, (d) assessment of possibility to provide users with suggestions for data improvement derived from information on the defective values.

ACKNOWLEDGMENTS

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

REFERENCES

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

Bicevska, Z., Bicevskis, J., Oditis, I. (2017). *Models of Data Quality*. In Information Technology for Management. Ongoing Research and Development (pp. 194-211). Springer, Cham.

Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2018a). *Data quality evaluation: a comparative analysis of company registers' open data in four European countries.* In Communication Papers of the Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 197-204).

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

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

Economist, T. (2017). *The world's most valuable resource is no longer oil, but data*. The Economist: New York, NY, USA.

Eppler, M. J. (2006). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes*. Springer Science & Business Media.

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

Gartner (2018). *How to Create a Business Case for Data Quality Improvement*. Available at: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/

Gartner (2013). *The State of Data Quality: Current Practices and Evolving Trends*. Available at: https://www.gartner.com/doc/2636315/state-data-quality-current-practices

Huang, K. T., Lee, Y. W., Wang, R. Y. (1999). *Quality information and knowledge management.* Publisher: Prentice Hall.

Nikiforova, A. (2018a). *Open 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 Baltic DB&IS 2018 Joint Proceedings of the Conference Forum and Doctoral Consortium, Trakai, Lithuania (Vol. 2158, pp. 00742158-1).

Redman, T. C., Blanton, A. (1997). *Data quality for the information age.* Artech House, Inc.

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.

The Register of Enterprises of the Republic of Latvia (2019). Available at: http://dati.ur.gov.lv/register/ (Accessed: 20 March 2019)

Company House (2018). Available at: https://www.gov.uk/government/organisations/companies-house (Accessed: 20 March 2019)

TDQM, (2019). T*he MIT Total Data Quality Management Program*. Available at: http://web.mit.edu/tdqm/ (Accessed: 20 March 2019)

DAMA, (2019). *The six primary dimensions for data quality assessment. Defining Data Quality Dimensions*. DAMA UK Working Group.

OpenDataSoft (2019). Available at: https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/ (Accessed: 20 March 2019).