D. Demmler, D. Krupka, H. Federrath. (Hrsg.): INFORMATIK 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 1023

Approaches for Automated Data Quality Analysis: Syntactic and Semantic Assessment

Agbodzea Pascal Ahiagble [1] and Hannah Stein[2]

Abstract: Data quality significantly influences data usability and plays an important role in data trading. This paper presents a data quality analysis (DQA) of data tables on two levels. The first, the so-called syntactic level, concerns the structure of the elements within the database and the second, the so-called semantic level, concerns the relationship between the elements in the database and the "real world". Based on a literature review the most relevant data quality criteria and corresponding metrics were derived. Subsequently, based on heuristics, a data-centric approach and an unsupervised machine learning clustering algorithm DBSCAN, a service for automated DQA, is designed and implemented (syntactic DQA). In the next step, an automated semantic DQA service as well. The approach is used to examine data tables for example for missing relevant columns (i.e., semantic completeness). A data quality index represents the services' output, which is derived from the automated analysis of various data quality criteria. This enables the assessment of data quality, as well as the detection of potentials for improving quality and thus increasing the value of tradeable data.

Keywords: Data quality assessment, data quality metrics, automated assessment services.

1 Introduction

Data play a very important role for companies' and institutions' activities, e.g., in the management of business processes. High-quality data bring benefits such as increasing customer satisfaction and competitive advantages, guaranteeing higher turnovers for the company [Pip02]. However, data with poor quality can have a negative economic impact on the company [Wan02], [STR96]. Companies and organizations, which work with large amounts of data, require approaches to automatically determine data quality (DQ) across different criteria, as this can be quite a time-consuming process. In data markets and ecosystems, clear responsibilities for DQ assessment are often lacking as well as automatic processes or clear control mechanisms to do so [Cap20], [Lis20]. In addition, data consumers cannot assess DQ before buying the data [Mus12] – a phenomenon that

[1] German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3, 66123 Saarbruecken, Germany, agbodzea_pascal.ahiagble@dfki.de
[2] Saarland University, Campus A5.4, 66123 Saarbruecken, Germany, hannah.stein@iss.uni-saarland.de

doi:10.18420/inf2022_85

is known as the information paradox [Sta17]. In this work, we present two approaches for decentralized, automatic DQ assessment, appropriate for future application in the context of data ecosystems, i.e., DQ assessment from a syntactic and semantic perspective. The services automatically determine the quality of databases that are generated and used in B2B and B2C contexts. We thereby mainly focus on the quality assessment of databases, defined as a "collection of all stored data and its associated descriptions" [Est00]. The quality of a database is derived from the quality of data or records collected in it.

Next, we present related work and elaborate on the most important DQ criteria for syntactic and semantic DQ assessment. Chapter three shows a methodology approach for developing the automatic assessment services. The technical conceptualization and prototypical implementation are described in chapter four. Chapter five concludes the paper.

2 Related Work

The definition of the quality of a database (DB) is considered from different points of view, e.g., quality is seen as "usability" [Eve07]. In this sense, the determination of quality depends very much on the data consumer and must be contextual [STR96]. The assessment of data quality can be made purely objectively [Lee06], which corresponds to the internal state of the DBs, or subjectively, what exactly the data consumers need or have experienced [Ang12], [Bal98], [Pip02], [STR96]. For this work, we have identified the most important DQ criteria, to be taken into account for semantic and syntactic DQ assessment.

2.1 Data Quality Criteria

The DQ criteria have been intensively researched over the last years and are determined independently of each other. The large number of criteria listed in the literature shows how multifaceted the assessment of data quality can be. Through research and analysis, it is possible to group these criteria, which can vary depending on the context. For example, [STR96] grouped the criteria into four categories, namely intrinsic DQ, accessibility DQ, contextual DQ, and representative DQ. Similarly, [Cai15] have grouped the criteria into availability, usability, reliability, relevance and presentation quality. Criteria such as accuracy, timeliness, reliability, completeness, relevance, accessibility, and interpretability are often used [STR96]. It is significant to know that data type has an impact on the DQ criteria as well as the evaluation and improvement technique [Bat09]. To evaluate DQ by criteria, metrics are used so that they can be represented as a value [Lev95]. In this paper, we focus on the DQ evaluation of structured data to enable automatic assessment service.

2.2 Syntactic Data Quality Assessment

"Syntactic data quality is concerned with the structure of data" [Sha99]. For [STR96] "data quality" is defined as data that are fit for use by data consumers. We have distinguished between evaluable criteria based on the internal elements in the database (objective) and evaluable criteria based on the external elements of the database (subjective). The first group includes all those that are measurable based on the variables provided directly by the database. We provide a summary which is developed based on a literature review in table 1. We name each DQ criterion, give its definition based on literature and then provide an example of how the criterion can be assessed.

Criterion	**Definition**	**Example**
Completeness	"The extent to which data are of sufficient breadth, depth, and scope for the task at hand" [STR96].	Missing values of a DB. Having "NULL" in a field instead of real value.
Integrity	"A measure of the existence, validity, structure, content, and other fundamental properties of the data" [McG08].	Entity integrity, referential integrity, domain integrity, and column integrity reflect only the states of the data [Cod90]
Accuracy	"The closeness between a value v and a value v' that is considered to be the correct representation of the real phenomenon that v is supposed to represent" [Bat06].	In a customer database where a customer has the first name Peter, "v = Peter" is correct and "v = pter" is incorrect.
Timeliness	“The extent to which the age of the data is appropriate for the task at hand" [STR96].	Delay between a change in state of the real world and the resulting change in state of an information system [Bat09], [Wan02].
Consistency	"The data is always presented in the same format and is compatible with the previous data" [Bat06].	Mismatch of datatype (numeric and alphanumeric) as attribute for the same entity.
Free-of-error	In general, this indicates whether "the data are correct" [Lee06].	The extent to which data are correct and reliable. [Pip02]

Table 1: Evaluable criteria based on the internal elements

Regarding the second group of criteria, with the inclusion of "external elements", the fact that the quality depends not only on the internal state of the data, but also on the use and satisfaction that the data bring is considered. Table 2 summarizes these DQ criteria using the same structure as table 1.

Criterion	Definition	Example
Credibility	"The extent to which the data is believed to be true and reliable" [Lee06].	Customers prefer to consume things they believe in without hesitation.
Interpretability and understandability	"The extent to which data are in appropriate languages, symbols, and units, and the definitions are clear" [Pip02].	Customers need to understand the data in order to use it in an appropriate way.
Security	“The extent to which access to the data is appropriately restricted to ensure its safety" [Pip02].	Data protection and confidentiality.
Representativeness	"The extent to which data is represented compactly and in the same format" [Pip02].	"Data are clear, without ambiguity, and easily understood" [Bat06].
Objectivity, Relevance, and Reputation	Objectivity is defined as "the extent to which data are unbiased and unprejudiced", relevance is defined as "the extent to which data are applicable and helpful to the task at hand" and reputation is defined as "the extent to which data are highly regarded in terms of their source or content" [Pip02].	They give data consumers an indication of whether the available datasets meet the baseline for their business processes.

Table 2: Evaluable criteria based on the external elements

2.3 Semantic Data Quality Assessment

Data quality can be defined from the semantic point of view as the "correspondence between the data represented by an information system and the data in the real world" [Ken98]. The criteria for semantic evaluation are similar to those for syntactic evaluation but with different measurement approaches, e.g., completeness, consistency, accuracy, and timeliness [Bat16], [Red98], [Yuv17]. In this context, a database is complete if it contains all the necessary information for the task to be performed [Jar20]. Furthermore, completeness can be defined as the degree to which all relevant attributes of a feature have been encoded [VER99]. Semantic consistency has been defined as the degree to which the evaluated data is free of internal inconsistencies [Ber18]. If we imagine a customer DB of a supermarket, for example, it is not semantically complete if the contact data of customers are not available as an entity. In the case that this characteristic

is present but requests are missing, we speak of syntactic incompleteness.

3 Assessment Approach

The assessment of DQ is specified by metrics. For the syntactic assessment, these metrics can be composed of simple to very complex formulas depending on the variables that come into play. For the construction of the individual metrics, [Pip02] use the simple ratio as a basic method. The quality index (QI) is thereby automatically measured and normalized as a value between 0 (very bad) and 1 (perfect) [Lee06].

According to [Bec08] there are two questions to ask when it comes to the semantic assessment of data quality: First, which real world phenomena should be represented in the information system (i.e., which entities are relevant)? For this purpose, a conceptual model must be created to capture through an ontology: the semantics of a real world. Second, how should relevant phenomena be represented in the information system (i.e., which attributes and relationships are relevant)? Therefore, a conceptual model is compared to the physical data model of an information system.

3.1 Automation of the Quality Assessment of Databases

In order to avoid redundant procedures and save time, it makes more sense to be able to automatically determine the quality of a database, especially now that companies and organizations have to deal with large amounts of data. Companies should be able to keep track of the quality of their databases without much effort, to be able to sell or share parts of them later via data markets and ecosystems. For example, [Bal85] designed a model to detect errors in data-based systems. The provision of data and how the relationships between quality parameters are represented should not be a challenge for companies [Wan02]. Strategies and techniques for assessing and optimizing DQ have been developed in recent years. According to [Bat09], the assessment of DQ is composed of three steps: (1) reconstructing the state of data with the goal of deriving general contextual information, (2) measuring the quality of the database, and (3) searching for improvement strategies and techniques. However, these steps may differ in practice, e.g. Evaluation of the DQ after the measurement or using iterative approaches.

Automation occurs on different levels. Activities that are not performed by humans but by the system can be described as automatic. In DQ assessment, this can be applied in the calculation of the metrics. Appendix of the available data sets, the relevant values for each metric can be collected automatically to perform calculations and return the score as a result. If values cannot be found, a warning of errors is issued instead of the QI as the result. Using heuristic functions [Sch18], [Kri17], targeted, statistical scores can be retrieved from one or more columns in the dataset. Through incremental computation, it is possible to include growing datasets when determining metrics. Machine learning allows, for example, using dataset and column names, to make a prediction about the

types of information present and thus, in case of discrepancies, can list them as errors. Training models with machine learning not only allow to find and fix specific errors such as lower and upper case to fully reflect the quality score [Sch18], [Ruk20], but also allows to independently and continuously monitor the quality level of the DB [Kri17].

4 Proof of Concept

To concretize the points described above, the implementation of a prototype is presented. This prototype is designed to automatically check both the semantic and syntactic quality of databases based on different criteria. According to the automation stages of [She05], our proposal can be classified to stage (2), since the user's intervention is required for the proposed prototype only when providing the records. The architecture of the prototype is illustrated in Figure 1 and has five components. All analysis is done with Pandas[3] and with Danfo.js[4] and visualization is done by Chart.js. As DB, only CSV files are considered here. The evaluation is a decentralized analysis, so it works on edge. The approach used for syntactic evaluation is an adaptation of the approaches proposed in [Pip02], [Sch18]. For consistency, the codes are written in Python.

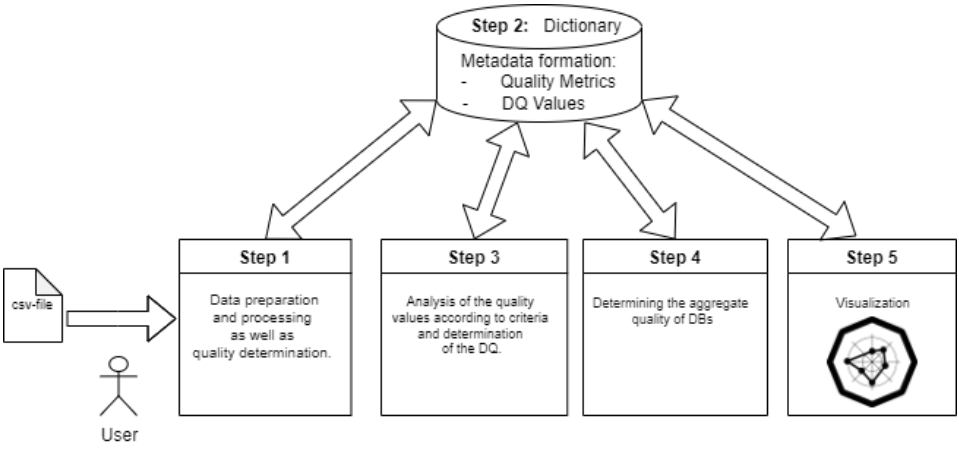

Figure 1: Prototype Architecture[5]

4.1 Preparation of Data Set

As mentioned above, the training data is CSV data. Here, care must be taken regarding

[3] https://pandas.pydata.org/
[4] https://danfo.jsdata.org/
[5] Include for syntactic evaluation step 1, 2, 3 and 4 while including step 1, 2, 3 and 5 for semantic evaluation.

separations (",", "\t", ";", "|"...) in the DBs. The first training data file is a small DB with 17 attributes and 1000 unique values from the platform Kaggle[6]. This dataset is a supermarket data recorded from 3 different shops over a period of 3 months. The second record has 73 attributes and 2686 rows and originates from a manufacturer. After uploading, the records (the CSV files) are stored as metadata in a dictionary[7] in the system, associated with an ID number. The conversion of the CSV file is done by using Pandas. First, we read the file as follows:

```
csv = pd.read_csv(csv_file, sep=delimiter)
```

Then we store the tables in the dictionary in the system to be able to continue working with them, as the code line shows:

```
data = csv.to.dict
```

4.2 Assessment of Quality Values

Pandas is used in different ways depending on the criterion to perform the evaluation. For syntactic quality assessment, four criteria are specifically included, namely completeness, consistency, accuracy and integrity of IDs [Ehr19]. This represents a data-centric approach. Additional inputs are required to consider other criteria [Lee06]. For Semantic Quality Assessment, completeness and consistency are also considered as criteria here [Ber18].

Syntactic Completeness

Pandas examines the dataset for empty entries, also called "Null Values". The number of "Null Values" is incrementally enumerated and then divided by the size of the dataset. Finally, the result of the division is normalized according to the "Simple Ratio" formula. The whole process looks like the following code in Python:

```
null_counter = int(csv[col].isnull().sum())
incomplete_cnt += val['null_counter']
metrics['completeness'] = 1-(incomplete_cnt/csv.size)
```

Syntactic Consistency and Accuracy

Here, mainly data type inconsistencies are considered. To find inconsistent values and inaccuracies, a heuristic function and a data-centric approach are used. As assumption, our method distinguishes between two types of inconsistencies, namely the so-called

[6] https://www.kaggle.com/aungpyaeap/supermarket-sales, accessed : 12.02.2022.
[7] https://realpython.com/python-dicts/. accessed 02/2021.

data type_inconsistency and the outliers per expression [Bat06]. The first type of inconsistency is composed of alternations of types within the same column. For accuracy, this is taken as the only violation characteristic. The second type, the outliers can be data entry errors, valid data but from a different population, or very rarely occurring data from the same population. As a basis of analysis, a heuristic based on statistical techniques is used: (1) The majority of values per expression are correct, with inconsistent values occurring infrequently and (2) errors can be different but equally distributed, or they can be the same but represent little compared to the correct entries. In order for this heuristic to be applied, a distinction was made between continuous and categorical data by inference. For the clustering of the data type DBSCAN[8] is used.

Semantic Completeness

For this category of evaluation, the absence of certain groups of attributes are considered as a violation [Jar20]. To overcome the challenge of a standard of minimum attributes that must be present in a database, database models of CRM vendors were analyzed and some interviews with different vendors were conducted. A database in the B2B domain should contain at least demographics, firmographics, technological, chronographic quantitative and qualitative data[9]. In this work master data in the B2C area were assumed as name, gender, address, email address and city. Within the database, these attributes are searched. In order to not miss any information, we first search for the synonyms of the attributes, which we include in the further processing as follows:

```
mydic = dict()
for i in basedata[10]:
    for val in wornet.synsets(i):
          for j in val.lemmas():
                array_synonyms.append(i.name())
```

Based on the number of master data and the number of matching attributes found in the DB, the semantic completeness is determined as follows:

```
Size_base = len(basedata)
semCompleteness = count[11] / size_base
```

The completeness decreases depending on how the attributes are available.

Semantic Consistency

The consistency on location, email address, gender and age are checked here. When

[8] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html, accessed 04/07/2022.
[9] https://www.cognism.com/b2b-data, accessed 04.07.2022.
[10] The master data collection
[11] Number of attributes occurring

checking the location data, the aim is to ensure that the specified location in the database actually exists in the real world. Therefore, individual locations are checked for existence in real-time using the free geographic database platform GeoNames API[12]. The code section for this looks like this:

```
Check_loc=requests.requests("GET",f"https://www.geona
                mes.org/search.html?q={Ci[13]}&country=")
```

Email address consistency was realized using the Python Library email-validator[14].

```
Regex = '^[a-z0-9]+[\._]?[a-z0-9]+[@]\w+[.]\w{2,3}$'
    If True == validate_email(m[15]) or
                (re.search(regex,m) == True):
                 isvalid.append(m)
```

For gender information, two gender types were considered as the base gender. The two genders are set in different formats and in three languages, namely English, German and French as default and searched in the database. We summarise it as code as follows:

```
Standard_Gender = ['female', 'male', 'weiblich',
          'männlich','feminin','masculin','f', 'm','w']
```

For age, all data above 100 and below 0 are considered as outliers as shown below:

```
If age < 100 and age > 0:
            not_outliers.append(age)
```

These four criteria are adjusted for the Quality Index and Simple Ratio.

4.3 Visualization

The ratings are presented as an index. The algorithm receives all uploaded data sets with their matching criteria as well as metrics as input. The whole is presented in the form of a radar, where the radius is equal to 1, which in turn corresponds to the QI of the "Simple Ratio = [0, 1]". Figure 2 shows how the visualization is presented

[12] http://www.geonames.org/, accessed : 05.07.2022.
[13] Current city
[14] https://pypi.org/project/email-validator/, accessed : 05.07.2021
[15] Current email address

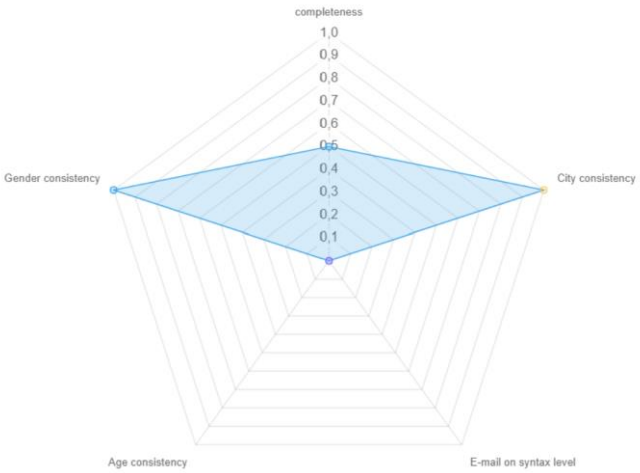

Figure 2: Visualisation of the supermarkt.csv evaluation

On the graph we can see that City Consistency and Gender Consistency have a QI equal to 1 (very good), which means that the data in the DB matches the data in the real world. So the cities really exist and the genders match the given standard. Age and e-mail consistency are equal to 0 in this case because they were not to be checked in the DB (see point: 3 description Fig.4). This leads to the QI of completeness being equal to 0.5 because this DB does not have a "Name" attribute in addition to these two attributes (see point: 1 description Fig.4).

4.4 Result

The results of the Proof of Concept (PoC) look in our interface as shown in the figures provided below. For the syntactic evaluation, the QI of the dataset and the aggregated QI are displayed first. For each dataset, the criteria values are used to determine how the QI was obtained. For the semantic evaluation, the QI per criteria is determined on the left side of the interface and, if necessary, the appropriate improvement options are indicated on the right side.

Record	Quality result	Data preview
Real life database.csv	88.20	Details
Average quality	**88.20**	

The quality criteria

Criteria	Values
Accuracy	0.9996327991921582
Consistency	0.9996327991921582
Completeness	0.5417843919256622
Id_integrity	0.9868215710074563

Figure 3: Syntactic evaluation of Real_life_database.csv

- The semantic completeness is: 0.5
- The semantic consistency in terms of locality is: 1.0
- The quality index of the e-mail at the syntax level is: 0.0
- The quality index of e-mail at the semantic level is: 0.0
- The quality index of gender at the semantic level is: 1.0
- The quality index of age at the semantic level is: 0.0

1. For semantic completeness, we recommend that you update your database with the following attributes:
['Name', 'Address', 'E-mail address']
2. The following given places are not to be found:
[]
3. For semantic consistency, we recommend that you check and update the completed email address:
['No email address was to be verified']
4. For semantic consistency, we recommend that you check and update the completed email address:
['No email address was to be verified']

Figure 4: Semantic evaluation of supermarkt.csv

5 Conclusion

Our work comes not without limitations which need to be overcome by future work. First, only four criteria were considered for the syntactic evaluation. Criteria like timeliness and accessibility, which are important for the companies, were left out. The inclusion of such criteria requires temporal information. Assuming that this information is provided by an external system, machine learning can be used in further assessment areas to extend the proposed model to include these criteria. For semantic evaluation, there is currently very little source to be found in the literature. Our PoC focuses on the DQ assessment of customer databases and is based on the analysis of customer database providers[16] [17] [18]. Further research in the context of semantic DQ assessment needs to consider further domains and types of data. In addition, we worked exclusively with structured data. Semi-structured data such as JSON files as well as unstructured data such as text or conversations, on the other hand, are difficult to evaluate without first adapting them [Blu03], [Tei16], [Ehr19]. An automatic procedure for evaluating this type of data would be important in future work.

In this work, we developed a prototype consisting of two services: one for syntactic and one for semantic DQ assessment. For the PoC, an open source database from the platform Kaggle[19] and a "real life" database from a manufacturer were used. Especially the semantic approach shows opportunities for more efficient and transparent DQ assessment to be applied for data sharing in data ecosystems.

6 Acknowledgement

This work is part-funded by the research project "Future Data Assets" (grant number: 01MD19010C) funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the "Smart Data Economy" technology program, managed by the DLR project management agency.

References

[Ang12] Angeles, M. d. P.; García-Ugalde, F. J.: Subjective Assessment of Data Quality considering their Interdependencies and Relevance according to the Type of Information Systems, International Journal on Advances in Software, Bd. 5, Nr. 3&4, pp. 389-400,

[16] https://www.oracle.com/webfolder/s/quicktours/cx/pt-customer-data-mgmt/index.html, accessed: 11.2021.
[17] https://help.salesforce.com/s/, accessed: 11.2021.
[18] https://www.teamleader.eu/blog/customer-data-management-how-to, accessed: 01.2022.
[19] https://www.kaggle.com/aungpyaeap/supermarket-sales, accessed : 12.02.2022.

2012.

[Bal85] Ballou, D.; Pazer, H.: Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems, Management Science, Bd. 31, Nr. 2, pp. 150-162, 1985.

[Bal98] Ballou, D.; Wang, R.; Pazer, H.; Tayi, G. K.: Modeling Information Manufacturing Systems to Determine Information, Management Science, Bd. 4, Nr. 44, pp. 462-484, 1998.

[Bat06] Batini, C. ; Scannapieca, M.: Data Quality Concepts, Methodologies and Techniques, Springer-Verlag Berlin Heidelberg, 2006.

[Bat09] Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A.: Methodologies for Data Quality Assessment and Improvement, ACM Computing Surveys, Bd. 41, Nr. 3, pp. 16:1-16:49, 2009.

[Bat16] Batini, C.; Scannapieca, M.: Data Quality Dimensions, Data and Information Quality, pp. 11-51, 2016.

[Bec08] Becker, J.; Matzner, M.; Mueller, O.; Winkelmann, A.: Towards a Semantic Data Quality Management using Ontologies to Assess Master Data Quality in Retailing, Proceedings of the Fourteenth Americas Conference on Information Systems, pp. 1-11, 14-17 August 2008.

[Ber18] Bernd, H.; Klier, M. Schiller, A.; Wagner, G.: Assessing Data Quality – A Probability-based Metric for Semantic Consistency.," Decision Support Systems, pp. 95-106, 2018.

[Blu03] Blumberg, R.; Atre, S.; The Problem with Unstructured Data, DM Review, 2003.

[Cai15] Cai, L. ; Zhu, Y.: The Challenges of Data Quality and Data Quality and Data Quality Assessment in the Big Data Era, Data Science Journal, Bd. 14, Nr. 2, pp. 1-10, 2015.

[Cap20] Capiello, C.; Gal, A.; Jarke M.; Rehof, J.: Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391), Report from Dagstuhl Seminar 19391, pp. 66 - 134, 2020.

[Cod90] Codd, E. F.: The Relatinal Model for Database Management, Eddison-Wesley, 1990.

[Ehr19] Ehrlinger, L.; Rusz, E.; Wöß, W.: A Survey of Data Quality Measurment and Monitoring Tools, 2019.

[Est00] Ester, M.; Sander, J.: Knowledge Discovery in Databases, Springer-Verlag Berlin Heidelberg, 2000.

[Eve07] Even, A.; Shankaranarayanan, G.: Utility-Driven Assessment of Data QUality, The DATA BASE for Advances in Information Systems, Bd. 32, Nr. 2, pp. 75-91, 2007.

[Jar20] Jarwar, M.,C.; Chong, I.: Web Objects Based Contextual Data Quality Assessment Model for Semantic Data Application, Appl. Sci. 2190(10), pp. 1-33, 2020.

[Ken98] Orr, K.: Data quality and systems theory. Communications of the ACM, 41, pp. 66 – 71, 1998.

[Kri17] Krishnan, S.; Franklin, M. J.; Goldberg, K.; Wu, E.; BoostClean: Automated Error Detection and Repair for Machine learning, 2017.

[Lee06] Lee, Y. W. ; Pipino,L. L. ; Funk , J. D. ; Wang , R. Y. : Journey to Data Quality, The MIT Press, 2006.

[Lev95] Levitin, A.; Redman, T.; The notion of data and its quality demension, Information processing & Managemant, Bd. 30, Nr. 3, pp. 9-19, 1995.

[Lis20] Lis, D.; Otto, B.: Data Governance in Data Ecosystems – Insights from Organizations, AMCIS 2020 Proceedings, pp. 1- 10, 2020.

[McG08] McGilvray, D.; Executing Data Quality Projects: TEN STEPS to Quality Data and Trusted Information, Morgan Kaufmann, 2008.

[Mus12] Muschalle, A.; Stahl, F.; Löser, A.; Gottfried, V.: Pricing approaches for data markets., International workshop on business intelligence for the real-time enterprise, pp. 129–144, 2012.

[Pip02] Pipino, L. L.; Lee, Y. W.; Wang, R. Y. ; Data Quality Assessment, COMMUNICATIONS OF THE ACM, Bd. 45, Nr. 4, pp. 211-218, 2002.

[Red98] Redman, C.: The Impact of Poor Data Quality on theTypical Enterprise, COMMUNICATIONS OF THE ACM, Bd. 2, pp. 78 - 82, Februar 1998.

[Ruk20] Rukat, T.; Lange, D.; Schelter, S. ; Biessmann, F.: Towards Automated Data Quality Management for Machine Learning, Amazon Science, 2020.

[Sch18] Schelter, S.; Lange, D.; Schmidt, P.; Celikel, M.; Biessmann, F.; Grafberger, A.: Automating Large-Scale Data Quality Verification, Proceedings of the VLDB Endowment, Bd. 11, Nr. 12, pp. 1781-1792, 2018.

[Sha99] Shanks, G.; Corbitt, B.: Understanding Data Quality: Social and Cultural Aspects, Proc. 10 785 th Australasian Conference on Information Systems, 1999, pp. 785-797, 1999.

[She05] Sheridan T. B.; Parasuraman, R.: Human-Automation Interaction, Reviews of Human Factors and Ergonomics, Bd. 1, Nr. 1, pp. 89-129, 2005.

[Sta17] Stahl, F.; Schomm, F.; Vomfell, L.; Vossen, G.: Marketplaces for Digital Data: Quo Vadis?, Computer and Information Science, Bd. 4, Nr. 10, pp. 22-37, 2017.

[STR96] Wang, R. Y., & Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information System, 12(4), pp. 5-34, 1996.

[Tei16] Teinemaa, I.; Maggi, M.; Fracescomarino, C. D.: Predictive Business proces Monitoring with Structured and Unstructured Data, Business Process Management, Bd. 14, pp. 402-417, 2016.

[VER99] VEREGIN, H.: Data quality parameters, Geographical information systems, pp. 177 - 189, 1999.

[Wan02] Wand, Y. ; Wang , R. Y.: Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), pp. 86-95, 1996.

[Yuv17] Yuval Z.; Even, A.:Development and evaluation of a continuous-time Markov chain model for detecting and handling data currency declines, Decision Support Systems, pp. 82 - 93, Januar 2017.