SOME ASPECTS OF DATA QUALITY PROVIDING

Ina Naydenova, Zlatinka Kovacheva

Abstract: *Providing quality data when working with information systems is an important and topical issue for many organizations. Poor data quality is often pointed as the source of operational confusions, inaccurate analytics and ill-conceived business strategies. It does not always lead only to a loss of money. The consequences could be much wider - creating public panic, loss of reputation or even loss of human lives. This paper presents some aspects of the data quality providing that are often misunderstood and therefore a significant setback in compiling and implementing a successful strategy for improving data. We also share some personal observations regarding the misunderstood aspects.*

Keywords: *data quality, context dependence, quality assessment, misunderstood aspects, process complexity.*

ITHEA Keywords: *E.Data, E.0 General, H.Information Systems, H.1.m Miscellaneous.*

DOI: https://doi.org/10.54521/ijita28-02-p01

Introduction

Increasing the amount of data that companies have used in recent years exacerbates their quality problems and the issue is becoming increasingly relevant. Nowadays, almost every organization has experienced the negatives of working with bad data. Examples of the economic damage that data quality issues can cause include added expenses when products are shipped to the wrong customer addresses, lost sales opportunities because of erroneous or incomplete customer records, and fines for improper financial or regulatory compliance reporting [Rouse et al. 2019]. It is estimated that correcting data errors and dealing with the business problems caused by bad data costs 15% to

25% of annual revenue on average for most companies [Redman 2017]. According to Gartner's Data Quality Market Survey, poor data quality costs organizations an average of $ 15 million in 2017 [Moore 2018].

However, poor data quality does not always lead only to a loss of money. The consequences can be much wider - creating public panic, loss of institutional reputation or even loss of human lives. In May 2020, the Australian government overestimated its spending commitments for a COVID 19 wage subsidy scheme by $ 39 billion, due to mistakes made filling in a confusing application form. The errors were missed and their aggregated value flowed into a bill passed by parliament [Farnworth 2020]. Also, the qualitative data is vitally important in healthcare, where people can be seriously affected as the result of incorrect information. Incorrect data at a Kaiser Permanante kidney transplant center caused deadly mismanagement of the donor waiting list [Redman 2008].

These are just a few of the examples that show why the provision of high data quality is a particularly topical issue and a matter of social and corporate interest. We believe that a good understanding of the process and nature of providing qualitative data is a key factor in both the development of a good strategy for improving data and its successful implementation. In this paper, we discuss some aspects of the data quality (DQ) that are often overlooked or misunderstood and hence a major obstacle to the success of the DQ programs. In addition, we share some personal observations regarding the discussed aspects in Bulgarian organizations.

What is Quality Data?

Firstly, we will focus on the question of what high-quality data means. According to many professionals, the data in an IT system is deemed of high quality if it correctly represents the real world. For example, if a customer's name is Julian Petrov, then his name should be recorded in the same way in the IT system. If there is a record Lian Petrov, then we have a quality issue.

However, in practice, the data is considered of high quality if it is good enough for the intended purposes of use. Some data is good ("high quality") for some purposes but is bad ("low quality") for others.

Example 1: If the customer's last name is used in a notification mail, there is no quality issue with the name Lian Petrov. The wrong first name, however, can be a problem if the person's identity needs to be verified.

Example 2: A business analyst assessing the yearly revenue of a product might consider the data about yearly sales of a product with seasonal variations (e.g. ice-cream) as quality one. The same data may not be good enough for a warehouse manager who is trying to estimate the orders for next month [Bertossi et al. 2011].

The examples suggest that quality always depends on the context in which data is used, leading to the conclusion that there is no absolute valid quality benchmark [BARC 2020]. Back in 1996, the Total Data Quality Management (TDQM) group of MIT University led by Professor Richard Wang has done in-depth research in the data quality area. Their observations show that many companies are improving data quality with practical approaches, but their improvement efforts tend to focus narrowly on accuracy. The researchers believe that data consumers have a much broader data quality conceptualization than information system professionals realize [Wang and Strong 1996]. They defined "data quality" as "fitness for use" and proposed that data quality judgment depends on data consumers. At the same time, they defined a "data quality dimension" as a set of data quality attributes that represent a single aspect or construct of data quality [Cai and Zhu 2015]. After the publication of [Wang and Strong 1996] work, many publications follow, and scholars propose different sets of dimensions. The majority of authors include some version of accuracy, completeness, consistency and timeliness among them [Sebastian-Coleman 2013].

Problems with data quality cannot be addressed effectively without an understanding of the data quality dimensions [Tayi and Ballow 1998]. Unfortunately, these dimensions are also context-dependent to some extent. Let us take, for example, the completeness: If we want to send notification letters we need a complete and accurate mail address, but if we are going to do marketing research based on the geographical areas, it is important only the city part of the address to be available. Therefore, depending on the specific

purpose for which we want to use a given set of data, it may be sufficient or unusable.

According to Wang and Strong, the dimensions can be grouped into four broad data quality classes: intrinsic, contextual, representational, and accessibility. Accuracy belongs to intrinsic, completeness, and timeliness to contextual and consistency to representational data quality classes [Wang and Strong 1996].

Table 1. Basic data quality dimensions

DQ Dimension	Meaning
Accuracy	The degree to which the data represents reality
Completeness	The degree to which required data is in the dataset
Consistency	The degree to which the data does not violate the semantic rules defined over a set of data items
Timeliness	The degree to which the data represents reality from the required point in time (i.e. how current the data is)

Context Dependence Challenges

- **Data Quality Assessment**

The discussed contextual nature of the quality is a root of many challenges. One of them is how to measure the severity of data quality issues in an organization effectively. Usually, people claim that they have a poor data quality based on their subjective experience. However, if the severity of the problem

cannot be assessed, it is difficult to convince the management to start data quality improvement initiatives.

Apart from assessing the severity of the quality issues, it is also difficult to estimate the financial benefits of providing better data. There are many reports mentioning millions in losses due to poor data quality (e.g. [Redman 2016], [Moore 2018], [Davie 2019]), but the calculation of a Return on Investment (ROI) is a complex task. Often the benefit is not only financial, but also reputational. It is quite embarrassing to tell a customer that you have no right to provide him with certain information because according to your system, he has passed away. The client's trust and his confidence in the quality of the provided services should also be taken into account in the profit formula. Generally, in producing estimates on costs caused by low quality data, it is difficult to identify all potential negative effects that are the result of low quality data, as well as all possible costs associated with assuring high quality data and their progression [Eppler and Helfert 2004].

The literature provides a wide range of techniques to assess the quality of data. Over time, these techniques have evolved to cope with the increasing complexity of data quality in interconnected information systems. Due to the diversity and complexity of the techniques, research has focused on defining methodologies that help select, customize, and apply data quality assessment and improvement techniques [Batini et al. 2009].

A comprehensive comparison between popular methodologies for data quality assessment and improvement is given in [Batini et al. 2009]. It is made with respect to different perspectives such as methodological phases and steps, strategies for assessing data quality, data quality dimensions and metrics, types of data (e.g. structured/unstructured), types of information systems addressed by each methodology and others. The comparison shows that methodologies tend to focus on a subset of DQ issues. Some methodologies focus on the assessment phase and provide limited support to the improvement phase, while others focus on the technical issues of both the assessment and improvement phases, but do not address economic issues. Although rare, complete methodologies are also available. They are helpful in providing a comprehensive framework to guide large DQ programs, but show the classical

tradeoff between the applicability of the methodology and the lack of personalization to specific application domains or technological contexts. Regardless of the focus, the definition of the quality dimensions and metrics to assess data is a critical activity in all methodologies. In general, multiple metrics can be associated with each quality dimension. In some cases, the metric is unique and the theoretical definition of a dimension coincides with the operational definition of the corresponding metric. In some methodologies, the data quality measurement is performed only with user questionnaires, while others use a mixture approach of questionnaires with a combination of subjective and objective metrics or with statistical analyses [Batini et al. 2009].

In [Eppler and Helfert 2004] is given a review and a classification of the potential data quality costs, either the costs of assuring data quality or the costs of low quality data. The review can help practitioners identify cost saving potentials and to get a holistic view of the aspects that could be taken into account when evaluating their data quality costs.

Our experience shows that very few organizations do data quality measurement. Usually, a set of observed problems are systematized and some examples are pointed based on a real-life experience. This is generally enough to convince the management that there are DQ issues and it is worth making an effort to solve them. Often, the launch of quality initiatives is motivated by regulatory requirements, which could lead to significant fines and financial losses for the company.

The profit from the quality improvement is also rarely calculated. The scope of the quality improvement project is determined according to the available budget and the stakeholders believe that the benefits will outweigh the money invested.

According to our observations, examples of data quality issues in operational and transactional systems that directly lead to reputational or financial losses are usually easier to be given. They have a very convincing effect on the senior management for the severity of the data quality issues. When it comes to the decision-support systems, it is more difficult to be pointed meaningful examples, although the extent of the DQ problems there is greater. Moreover, the wrong

corporate strategy or a marketing decision could cause much greater losses for the company than operational DQ issues could.

- **Data quality in Big Data solutions**

The contextual nature of data quality raises the question of usability of solutions that accumulate large amounts of data from various sources, without clear vision about their purposes of use (e.g. Big Data solutions). How to ensure big data quality and how to analyze and extract knowledge hidden behind the data become major issues for industry and academia in the last decade [Cai and Zhu 2015]. Poor data quality will lead to low data utilization efficiency and even bring serious decision-making mistakes. According to [Reid et al. 2015] two-thirds of European and American businesses are not able to unlock value from Big Data. A more recent study found similar results. New Zealand's Massey University survey shows that nearly two of three managers said they had no confidence or trust in Big Data, preferring to rely instead on their intuition and experience to make decisions [Massey Uni 2019].

An extensive and systematic literature review in the Big Data quality area is given in [Mirzaie et al. 2019]. The authors analyze, compare and classify numerous studies, looking for answers to various questions, such as what the core research topics are, what kind of methods and technologies are proposed, what are the main open challenges in the field and others. The analyzed studies discuss various techniques for detection of bad data (mainly outlier detection techniques), methods for data quality evaluation and cleaning algorithms. The techniques that are used for bad data detection and cleaning are very similar. Most of them are based on machine learning algorithms and statistical models. In contrast, the evaluation methods are based on more traditional approaches, such as the examination of the extent to which the data items satisfy different types of constraints defined on them. Depending on whether the studies focus on batch (offline/ static) or stream (online) processing, different algorithms and approaches are used. In addition, hybrid approaches have become increasingly popular in recent years. These approaches work both on batch and stream data, typically using static data to build a model and apply the model to the data stream. The techniques presented in almost half of the articles are not specific to a particular domain. Mirzaie et al. also examine other publications that do a

review of the literature in the field of Big Data quality. Their examination shows that the reviews did interesting classifications, but they are not systematic enough.

Based on the large number of reviews, we can conclude that the interest in ensuring the quality of data in the Big Data area is high, but scattered. Also, considering the recent Big Data academia research topics and results of different surveys on the trust of the decision makers in their analytics (e.g. [KPMG 2017], [Massey Uni 2019]), it can be concluded that quality of the data, and hence its usability is still one of the most important issues facing Big Data applications. Although researchers have proposed different techniques to ensure quality, the research on Big Data quality is in its infancy stage [Zhang et al. 2017].

Our observations are that the research in the field of Big Data quality focuses on the use of specific techniques, as well as the challenges associated with large volumes of information, computational resources and the availability of unstructured data. However, the topic of using appropriate methodologies, describing the whole process to achieve better quality in Big Data solutions, particular steps and criteria for choosing one or another technique, including specifics arising from the usage context or specific data domains are rarely addressed. The contextual nature of quality suggests that it makes sense as a first step in the data quality provisioning to fix the usage purposes, and after that to consider the appropriate assessment and improvement techniques. Based on the available publications, we can conclude that the subjective nature of the data quality is often overlooked in the Big Data area. The expectation that objective data quality improvement necessarily implies business value is widespread [Loshin 2011]. This is understandable to some extent - since the data sources are external and cannot be managed, the "process-driven" approach, which relies on redesigning the processes that create or modify data, is not applicable here. However, the question of how much value can be derived and how cost-effective are evaluation and cleaning activities in the Big Data applications remains open. Our expectations are the studies in this area to evolve in a direction that ties the data quality metrics and techniques to the business's performance, moving from a data quality to an information quality.

- **"The more the better!" is not always true**

Aside from Big Data applications, many enterprises tend to store redundant information in their corporate data warehouses. Concerns that in future they may need data that is not quickly accessible often leads to the accumulation of significant amounts of redundant data. However, "The more the better!" is not always true. Over time, some of this data begins to be used by various players in the organization. They rarely are aware of the data accuracy and timeliness. This increases the issues with data duplication, data integrity and consistency in the organization. As a result, a lot of effort is put into clarifying the reasons for data discrepancy between various departments and in general, business users lose confidence in the quality of data they use. On the one hand, business analysts want flexible BI solutions, but on the other, they have reservations about using ad-hoc functionalities due to a lack of trust in data. They prefer to use predefined reports, because their data have been tested and reliable. Indeed, the lack of trust in data on the part of corporate executives and business managers is commonly cited among the chief impediments to using business intelligence (BI) and analytics tools to improve decision-making in organizations ([Rouse et al. 2019]).

Our experience shows that in recent years, organizations have begun to realize the cost of accumulating redundant information - it's not just about hardware costs, but also the cost of human labor in terms of management, data control, extra time and issues during the data loading, increased risks of security breaches, data quality issues etc. Nevertheless, many organizations still tend to clutter with data. There are still managers who are supporters of the maxim "Let us have the data and then will think how to use it". This indicates a certain immaturity and lack of awareness about the consequences of data hoarding approach.

Our experience also shows that General Data Protection Regulation (GDPR), which comes into force in 2018, has helped many organizations to realize the mentioned side effects and start thinking in a direction of providing master data management solutions. The regulation itself does not set limits on the amount of information collected, but requires a clear vision of what personal data is processed in the organization and for which purposes according to the defined

ones in regulation. If the organization cannot specify a lawful basis for processing a certain personal data, the data must be deleted.

Who is responsible to provide a Data Quality

The next aspect to discuss is who is responsible for data quality providing. The IT department is usually held responsible for maintaining quality data, but those entering the data are not ([Hickey 2016]). According to our experience, the misconception that data quality is mainly an IT problem and it could be solved by purchasing sophisticated (and often expensive) data quality tools is very common. This misunderstanding leads to projects with unfulfilled or partially achieved objectives.

One of the reasons for such delusion could be the manner in which the quality tools are advertised. The ads often suggest that scientific approaches embedded in the tools can easily solve any data issue. For this purpose, it is enough to have a professional with a good IT background to set up the quality improvement process. For many business specialists, the smart technologies provided by the tool, such as profiling, artificial intelligence, data-mining algorithms and so on, make their participation in the quality issues resolving unnecessary – the tool could solve the issues better than anyone.

Another possible reason for such delusion is the assumption that developers bear primary responsibility for ensuring that the software they write works properly no matter which data is fed into it. Since they write the code, they have the power to control how an application will respond when it receives low quality data. However, in the real world, the amount of time and effort that programmers devote to handling potential data quality problems in their software is quite limited. They usually include a code to perform only a basic data validation, which ensures that data input into an application is complete, formatted as expected etc. Moreover, even the best programmers cannot foresee every type of data quality issue that could occur within their applications. Even if they did, the applications might end up being overloaded by functions that handle obscure data quality issues. Therefore, it is hardly realistic to expect programmers to address every potential data quality issue that could impact their applications ([Tozzi 2020]).

Actually, the organization and its processes are always more important than the technology since they are defined according to company strategy ([BARC 2020]). The business professionals must not only be involved, but also lead the data quality initiatives. This is a consequence of the context nature of the quality. In fact, the businesses know the context and the indented uses of data better than anyone does.

In [Breur 2009] the author believes that data quality is, or should be, everyone's concern. To enable business alignment, it is necessary to bring together the problem holder and problem owner. The problem holder is the person who experiences "pain" from a problem; a problem owner is the person who controls the resources needed to resolve it. The authors give an example in which the IT specialists are in the role of holders: the team that is responsible for the extract, transformation and loading activities (ETL) in the corporate data warehouse is often facing conflicting data from disparate sources. In this case, system owners of supplying systems are problem owners: they own the data. A long-term problem solving usually requires structural changes and/or amendments in business processes of the supplying systems. The example illustrates that an all-important topic as the data quality cannot be left only to the IT department. In addition, the participation of people with different job positions in quality programs helps to build a right culture and attitude towards the DQ assurance process in an organization.

After all, what is the right approach of how to provide better data quality? Data quality experts from research and industry agree that a unified framework for data quality management should bring together organizational, architectural and computational approaches ([Sadiq 2013]). An important factor for the success of the program is the careful definition of the data domains and their DQ managers, participant roles and their responsibilities (e.g. problem holder and data owner, data managers, data stewards etc). The overall framework of all data quality and cleansing processes should be designed, including the specification of particular DQ processes, DQ activities and responsibilities in each step. There must be a readiness for changes in the established business processes. It is a good practice the procedures by which such changes will be made to be coordinated in advance. Last come the architectural and

computational approaches to support the implementation of the specified processes and to automate the DQ activities where it is possible.

Data Quality Life Cycle

Another common mistake is the data quality providing to be considered as a one-time initiative, rather than an ongoing process.

The discrepancies in the data should be detected, but also effectively rectified. Usually this cannot be achieved with one-time initiatives. The elimination of discrepancies must be done in a consistent manner, especially when there is a transfer of data streams in the organization. Every copy of the bad data should be corrected. This implies the use of a DQ maintenance program, which consists of clearly defined data correction procedures, clearly defined decision-making rights and constant monitoring.

The DQ program participants should be aware that "Everything flows and nothing abides". The business is constantly changing, and this requires appropriate changes in the quality data program, including a change in already adopted cleaning procedures, revision of the established business rules for data verification, and calibration of metrics for quality assessment and so on.

Of course, the one-time initiatives also have their place. They are suitable for errors that have arisen in batches or because of data migration procedures, but for process causes of data non-quality, a data quality program is more appropriate. Since both kinds of errors often occur side by side, and also to ensure both short-term improvement and sustainable success, in practice a combination of these two approaches is often called for ([Breur 2009]). According to a study conducted in ([Hickey 2016]), the main source of bad data in an organization is improper data entry by employees. Data migration or conversion projects come in second, followed by mixed entries by multiple users, changes to source systems, systems errors, data entry by customers and external data.

The effort required to maintain the quality of data use in a system that is in operation should also not be underestimated. Let us share some personal experience: we have implemented an IT system that consolidates data from

different companies at a holding level. We do a business rules validation on the input data in order to provide high quality of data. The detected problems are divided into two classes: "errors" and "warnings". Errors alert to a situation that is not acceptable and the wrong data cannot enter the system until the problem is fixed. Alerts indicate a potential data error or a problem that is not critical. Shortly after the system is put into operation, we have a request from data supplying systems owners to transform particular "errors" into "warnings", in order to allow certain information to be entered into the system. The reason for this request is that the supplying systems cannot provide accurate data. Our answer was that we would fulfill the request, but first we had to analyze how this action would affect the current algorithms, the established reports and KPIs, because they were written with the presumption that certain situations were unacceptable; if they became acceptable, this could break our logic. In other words, we need to check if the data will remain good enough for purposes, for which they are used so far. For a medium-sized system, this involves a lot of work. However, the change requestors were surprised how much work had to be done, because from their point of view this was only a simple change.

The given example illustrates that the price of life cycle changes is not cheap, but it should be borne in mind if the organization wants to maintain a good data quality over time.

Drowned in Complexity

Another problem that organizations face in providing quality data is the complexity of the project. Many companies are afraid of data optimization projects, because the organizational, procedural and technological adjustments that have to be made can seem too complex and incalculable ([BARC 2020]).

Our observations are that the main reason for such concerns is that many organizations underestimate the phase of "AS-IS" analysis. They avoid starting a separate project for "AS-IS" analysis and a next project for "TO-BE" strategy implementation. As a result, the allocated time and resources for the analysis phase are not enough. Many business professionals are not aware that they have to become part of such analysis, so they are not interested in the project aims. Other participants in turn focus on analyzing and profiling the data quality

state in the databases, instead of interviewing the business users about the problems encountered by them. The organization remains unable to see the big picture and it is not capable to decide how to achieve maximum benefit with minimum effort. The goal of the quality improvement program should not be to achieve a perfect data quality across the organization, which is indeed complicated or even an impossible task, the intelligent attitude implies to be looking for a reasonable improvement of the current state.

We see two basic reasons why company's managers avoid splitting the quality improvement program into separate projects. The first one is that the senior managers prefer to fix a budget for the whole project (both analysis and implementation). Postponing the decision of how much money to invest for the quality implementation based on the severity of revealed issues involves a more complex budgeting scheme, which is not preferred. The second one is that they consider as a risk to outsource the analysis to one contractor and the implementation to another. This is mainly because there are not many professionals who are able to perform and describe qualitatively the results of such kind of analysis. In addition, the companies themselves do not have sufficient experience in requiring the analysis outcomes to be presented in a way that allows the outcomes to be understood and used by everyone.

The methodologies for DQ improvement also reflect the underestimation of the need for a holistic analysis. Almost all methodologies include activities, which examine data schemes and perform interviews to reach a complete understanding of data, but only a few of them consider the data quality requirements an analysis step, identifying data quality issues from the business perspective and collecting new target quality levels [Batini et al. 2009]. Although it is reasonable the DQ issues and their degree of importance according to business users to be collected first, and after that the data schemes to be examined for assessment of the issues' extent, this is often not the case.

The quality of the initial analysis is one of the main factors for the success of the DQ program. The proper analysis allows drawing a reasonable long-term strategy to solve the most severe problems as a first step and gradually evolve to cover a wider range of problems. The practice shows that companies having successful data quality initiatives achieved their success because they took the

decision to launch their projects and then evolved step-by-step ([BARC 2020]). In addition, it should not be forgotten that the way to success includes compromises and a wise selection of what is worth to be improved, as there is no perfect data.

Acknowledgements

The paper is published with partial support by the Task 1.2.5 of the Bulgarian National Scientific Program "ICT in Science, Education and Security", funded by the Ministry of Education and Science (Contract MES DOI-205/23.11.2018)

Conclusion

The contextual nature of data quality complicates the process of quality providing and places many challenges. However, if people are aware of the data quality specifics, if they know where the pitfalls are and how to avoid them, then they have a very good chance to define an effective quality strategy and to implement it successfully. For this reason, we have considered the main aspects of data quality provision, which poor understanding could lead to difficulties in drawing up or implementation of DQ programs. We discussed the contextual nature of the data quality, some challenges regarding to data quality assessment and the storage of data without a clear vision of its use, the question of who is responsible for DQ providing, important accents of the DQ life cycle and the role of the initial analysis for the success of DQ program. Our belief is that an in-depth understanding of the aspects addressed in the paper is a significant factor in achieving productive data quality initiatives and will contribute to more and successful quality programs.

Bibliography

[BARC 2020] Business Application Research Center. Data Quality and Master Data Management: How to Improve Your Data Quality. Business Application Research Center Imprint, 2020. https://bi-survey.com/data-quality-master-data-management

[Batini et al. 2009] C. Batini, C. Cappiello, C. Francalanci, A. Maurino. Methodologies for Data Quality Assessment and Improvement. ACM

Computing Surveys, Vol. 41, No. 3, Article 16, https://www.researchgate.net/publication/220565749_Methodologies_for_Data_Quality_Assessment_and_Improvement

[Bertossi et al. 2011] L. Bertossi, F. Rizzolo, L. Jiang. Data Quality Is Context Dependent. BIRTE 2010, Lecture Notes in Business Information Processing, vol 84., pp 52-6, Springer, Berlin, Heidelberg, 2011. https://link.springer.com/chapter/10.1007/978-3-642-22970-1_5

[Breur 2009] T. Breur. Data quality is everyone's business - Managing information quality - Part 2. Journal of Direct, Data and Digital Marketing Practice volume 11, pages114–123 (2009). https://link.springer.com/article/10.1057/dddmp.2009.21

[Cai and Zhu 2015] L. Cai, Y. Zhu. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, p.2, 2015. https://datascience.codata.org/articles/10.5334/dsj-2015-002/

[Davie 2019] M. Davie. Why Bad Data Could Cost Entrepreneurs Millions. Entrepreneur Media Inc., 2019. https://www.entrepreneur.com/article/332238

[Eppler and Helfert 2004] M. Eppler, M. Helfert. A classification and analysis of data quality costs. Proceedings of the Ninth International Conference on Information Quality (ICIQ-04) 2004. https://www.researchgate.net/profile/Martin_Eppler/publication/266422881_A_classification_and_analysis_of_data_quality_costs/links/54e285680cf2c3e7d2d3a8e0/A-classification-and-analysis-of-data-quality-costs.pdf

[Farnworth 2020] R. Farnworth. The Six Dimensions of Data Quality - and how to deal with them. Towards Data Science, 2020. https://towardsdatascience.com/the-six-dimensions-of-data-quality-and-how-to-deal-with-them-bdcf9a3dba71

[Hickey 2016] N. Hickey. Data quality: Whose job is it?. GCN 1105 Media, 2016. https://gcn.com/articles/2016/07/13/data-quality-responsibility.aspx

[KPMG 2017] KPMG. Trusted analytics matter to CEOs. Trusted Analytics article series, August 2017, Issue 7. https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/08/da-trusted-analytics-ceo-outlook.pdf

[Loshin 2011] D. Loshin. The Practitioner's Guide to Data Quality Improvement. Book, ISBN: 9780123737175, 2011

[Massey Uni 2019] Massey University. Study shows many senior managers distrust big data. Science X Network, 2019. https://phys.org/news/2019-04-senior-distrust-big.html

[Mirzaie et al. 2019] M.Mirzaie, B.Behkamal, S.Paydar. State of the Art on the Quality of Big Data: A Systematic Literature Review and Classification Framework. Cornell University, 2019. https://arxiv.org/ftp/arxiv/papers/1904/1904.05353.pdf

[Moore 2018] S. Moore. How to Stop Data Quality Undermining Your Business. Gartner, 2018. https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business

[Redman 2008] T. Redman. Data Driven: Profiting from Your Most Important Business Asset. Harvard Business Press, 2008. https://books.google.bg/books/about/Data_Driven.html?id=Y0G6HpIMbVUC&redir_esc=y

[Redman 2016] T. Redman. Bad Data Costs the U.S. $3 Trillion Per Year. Harvard Business School Publishing, 2016. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

[Redman 2017] T. Redman. Seizing Opportunity in Data Quality. MIT Sloan Management Review, 2017. https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

[Reid et al. 2015] C. Reid, R. Petley, J. McClean, K. Jones, P. Ruck. Seizing the information advantage: How organizations can unlock value and insight from the information they hold. PricewaterhouseCoopers, 2015. https://www.pwc.es/es/publicaciones/tecnologia/assets/Seizing-The-Information-Advantage.pdf

[Rouse et al. 2019] M.Rouse, C.Stedman, J.Vaughan. What is data management and why is it important?, Last updated 2019. https://searchdatamanagement.techtarget.com/definition/data-quality

[Sadiq 2013] S.Sadiq. Handbook of Data Quality, Research and Practice. Book, ISBN:9783642362569,2013,https://www.researchgate.net/publication/320597717_Handbook_of_Data_Quality_Research_and_Practice

[Sebastian-Coleman 2013] L. Sebastian-Coleman. Measuring Data Quality for Ongoing Improvement. A volume in MK Series on Business Intelligence (2013). https://www.sciencedirect.com/book/9780123970336/measuring-data-quality-for-ongoing-improvement#book-info

[Tayi and Ballow 1998] G. Tayi, D. Ballou. Examining data quality. Communications of the ACM Vol. 41, No. 2, 1998. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.141.7860&rep=rep1&type=pdf

[Tozzi 2020] Ch.Tozzi. Solving Data Quality Problems Is Not (Only) Programmers' Responsibility. Precisely Blog, 2020. https://www.precisely.com/blog/data-quality/solving-data-quality-problems

[Wang and Strong 1996] R. Wang, D. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12:4(1996), pp.5-34. http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf

[Zhang et al. 2017] P. Zhang, X. Zhou, J. Gao, Ch. TaoQuality. Assurance Technologies of Big Data Applications: A Systematic Literature Review. IEEE BigDataService 2017 - International Workshop on Quality Assurance and Validation for Big Data applications, 2017. https://www.computer.org/csdl/pds/api/csdl/proceedings/download-article/12OmNzlD9CC/pdf

Authors' Information

Ina Naydenova *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences; Sofia, Bulgaria; e-mail: naydenova@gmail.com*

Major Fields of Scientific Research: Business Intelligence, Data Warehouses

Zlatinka Kovacheva *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences; University of Mining and Geology, Sofia, Bulgaria; e-mail: zkovacheva@math.bas.bg*

Major Fields of Scientific Research: Neural networks, Big data