OULUN AMMATTIKORKEAKOULU

Katri Mäkitalo

USING CUSTOM DATA QUALITY MATURITY TOOL FOR IMPROVING DATA QUALITY IN AN ORGANIZATION

USING CUSTOM DATA QUALITY MATURITY TOOL FOR IMPROVING DATA QUALITY IN AN ORGANIZATION

Katri Mäkitalo
Master Thesis
Autumn 2023
Degree Programme in Data Analytics
and Project Management
Oulu University of Applied Sciences

ABSTRACT

Oulu University of Applied Sciences
Degree Programme in Data Analytics and Project Management

Author: Katri Mäkitalo
Title of the thesis: Using custom data quality maturity tool for improving data quality in an organization.
Thesis examiner: Ilpo Virtanen
Term and year of thesis completion: Autumn 2023 Pages: 66

Data is money, and bad data is loss of money. Nowadays, huge amount of data is created, and the amount of data is expected to increase in the future. This data is used for traditional reporting and decision making, but also as a basis for automation, artificial intelligence, and machine learning. It can also be used for automated business operations, and even automated decision making. For these purposes, good quality of data must be ensured because the benefits of data-driven organization can only be achieved with good quality data.

This research was qualitative yet empirical with measurement done via observations in a case company while experimenting and piloting customized data measurement, using a data quality measurement tool. In addition, there was a literature review which covered various aspects of data quality and data quality maturity models. Both data quality dimensions and data quality maturity models were covered in the literature review. Additionally, concrete actions for the possibilities to improving data quality and data quality maturity were included in this thesis.

Results of this study showed that the customized data quality maturity assessment can help case company to improve data quality maturity in the long run. This study was too short for making the data quality improvements visible via the custom data quality maturity tool and actions based on the results of using said tool, but concrete steps to improve data quality and DQ management were indicated by the data quality maturity measurement tool to be achieved via concrete actions in the future. However, data quality awareness and understanding of data quality were improved in the case company during this research. There is also a plan to experiment using this custom data quality maturity tool for KPI status, target settings and continuous measurements.

Keywords: data quality, data quality improvement, maturity model, data quality dimensions, data quality solutions, data profiling, data quality measurement

CONTENTS

ABBREVIATIONS

AI	Artificial intelligence
ADQ	Augmented Data Quality
API	Application Programming Interface
CPI	Consumer Price Index
DQ	Data Quality
DQM	Data Quality Management
DLC	Data Life Cycle
DG	Data Governance
DM	Data Management
ESG	Environmental, Social, and Governance
ETA	Estimated Time of Arrival
ETL	Extract, Transform, Load
ELT	Extract, Load, Transform
EU	European Union
GDPR	General Data Protection Regulation
IOT	Internet of Things
IT	Information Technology
ML	Machine Learning
PII	Personal Identification Information (e.g., name, address, social security number, passport information, biometric records)

1 INTRODUCTION

The amount of collected and available data for organizations is rapidly increasing and it will continue to increase in the future (Statista, 2023). In addition to the traditional, internal data from enterprise systems, data is now available for all organizations across industries from many external sources and IOT systems. Having transparency and taking care of the data quality is, and will be, even more important now when data is not only used for traditional reporting and decision making, but also as a basis for automation, artificial intelligence, and machine learning. Then again, the results of all this information can be used for automated business operations and even automated decision making. Poor data quality can lead to misguided decisions, inefficient processes and decreased customer and employee experiences. Benefits of the efforts for being a data-driven organization can only be achieved with good quality data.

Data quality management has been studied and discussed since the 1980s (Zhu, Madnick, Lee, Wang., 2012) but concrete, systematic data quality improvement actions are now required more than ever in order to ensure the quality of data in every organization's operations and decision making. All the organizations should also ensure that the data quality of the data received from external systems meets the quality standards of their intended purposes. It is a must that the data quality is monitored in every step of the data flow, and this information is also available for everyone who uses it.

The fact is the data is much more expensive to correct afterwards. The more the data is validated already at the source systems, the less expensive it will be for the organization. George Labovitz and Yu Sang Chang proposed the 1:10:100 already in 1992 (Grepsr, 2021) which in this context means these costs for the data quality are like *"$1 for the prevention, $10 for correction, $100 for doing nothing*" (Grepsr, 2021). Today, on-premises and/or legacy IT systems with several traditionally developed and maintained integrations have created more complex IT system environments than in the '90's so much more dollars could easily be added to this rule. Although cloud-based systems have lately decreased that complexity, data created today may well end up used by many other systems, automations, and AI/ML developments, not just the system creating it or directly interfacing it.

Historical data is also collected for the reporting and ML/AI purposes and the errors in the data can also have an effect in the future. In some cases, it can be too expensive, or even not possible, to correct the errors in historical data afterwards, like for example the error by Statistics Finland where electricity prices were included twice in the Consumer Price Index due to problems in the data sources. This error could not be corrected retrospectively because of international principles. Because of this, the error effected the decisions based on Consumer Price Index, for example pension index and social benefits so it caused extra costs for the government. (Statistics Finland, 2023).

An even worse example is from the Netherlands tax authority which used a self-learning algorithm to find out benefits frauds. Thousands of people were affected over six years because of wrong labels in the algorithm before these errors in labels were found. Penalties, based on this implementation, took place based only on "*suspicion of fraud based on the system's risk indicators*". For over 6 years, people were often wrongly labelled as fraudsters and things like low income or having dual nationality and "a non-Western appearance" were marked as a big risk indicator. Consequences of that error was devastating for the effected people like poverty which also led to suicides by some victims and "*more than a thousand children were taken into foster care*". Tax authorities were fined by €3.7 million by the country's privacy regulator based on EU's data protection rulebook (GDPR, 2012) for example on the basis that they did not have a legal right to process the data and they also kept the data and information in their databases for too long (Politico, 2022).

On the other hand, innovative businesses use good quality data to drive growth and transform themselves. According to Gartner research, Airbnb and Amazon are good examples of the companies who use good quality data to take advantage of all their data assets and accelerate growth (Gartner, 2018).

> "*Good quality data empowers business insights and starts new business models in every industry. It allows enterprises to generate revenue by trading data as a valuable asset*." (Gartner, 2018)

For example, Airbnb validates product ideas by using "*randomized controlled experiments*" and track their business performance rigorously for maximizing their value for stakeholders (Chang, 2021). Airbnb has invested in concrete data quality development actions for many years already, for example by creating and using their "*Midas Certification Process*" for certifying their data assets (Quoss, 2020). That has brought "*a dramatic increase in data quality and timeliness to Airbnb's most critical data*" (Wright, 2023). However, they have continued their data quality development

actions according to the experiences which they have got from using “Midas certification”, and have enhanced and scaled their data quality processes and practices further from this. They have now decided to rely on incentivization of both the data producer and consumer, and introduced a data quality score which is tied to data asset (Wright, 2023).

Without concrete actions, data quality will not be improved. To set targets and follow up these actions systematically, measurements are also needed for the data quality maturity. The data quality maturity of each business-critical data asset should be transparent to the management and the entire organization. There will also be regulations like EU AI Act which will be in force in near future and these regulations include “data quality requirements for inclusive, non-biased and trustworthy AI” (European Commission, Joint Research Centre, Balahur, Jenet, Hupont Torres et al, 2022). The question is, how to achieve all these steps which are needed to improve data quality, get the current state of the actions, and follow up the status of these improvement actions? Holistic approach for using both data quality and custom data quality maturity measurements to improve data quality and getting transparency of the data quality improvement actions are included in this study (Figure 1).

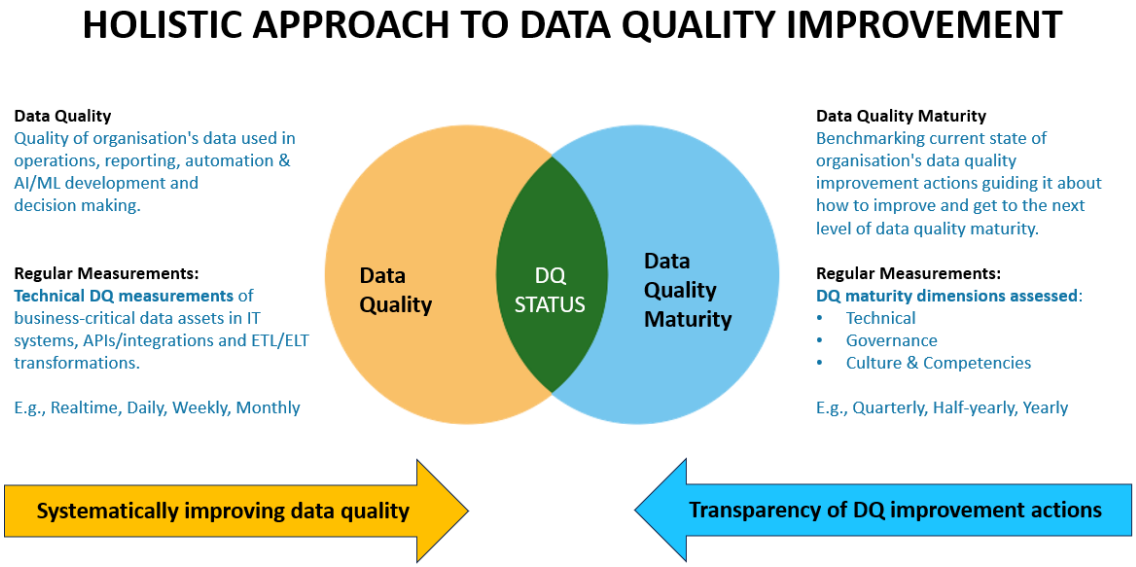


Figure 1. Holistic approach to data quality improvement for systematically improving data quality and getting transparency of the data quality improvement actions. Holistic data quality status is illustrated in green colour.

1.1 Research objectives and questions

The objectives of this study were to experiment and observe the usage of customized data quality maturity template for assessing the data quality maturity in a case company for selected business-critical data assets in different business lines. In addition, target was to give an overview of the different aspects related to data quality like data quality dimensions, effects of poor data quality and data quality improvement possibilities. For these purposes, publicly available, free to use data quality and data quality maturity model related literature was used. Via the literature reviews it was noted that typically either the data quality dimensions or data quality maturity models are included in the research but because the data quality maturity is dependent on the quality of the DQ dimensions, both the basics of data quality dimensions and data quality maturity models were included in this thesis.

Target is to contribute to the academic discourse on data quality and DQ maturity models, bridging the gap between theoretical models and practical implementation. The hypothesis of this study is that the usage of this kind of custom data quality maturity assessment tool helps the organization to acquire more awareness and understanding of the data quality in general. In addition, to also get transparency to the current state of the data quality maturity especially of the business-critical data assets. Eventually, by gradually improving the data quality awareness and understanding in an organization, and in case correctly incentivized, systematic and concrete data quality improvement actions can be increased. Then it would also be possible to include these data quality improvement actions into the business processes. These improvement actions can be regularly measured and reported by using the custom data quality maturity assessment tool.

The research questions were:

1) Is this kind of custom data quality maturity assessment tool useful in improving data quality of an organization in a systematic way?
2) How does this custom data quality maturity assessment tool compare to the existing publicly available, free to use data quality maturity models?

1.2 Research methods

The research was qualitative data collection done via self-report and observations. It entailed experimenting and piloting customized maturity template created in MS Excel as a heatmap and then using that in data quality training session in group work discussions and self-report by each group. There were five groups and each group had 4-6 participants from different business lines. As a result, heatmap of self-evaluations was updated by each group. Group work exercise was also done for creating awareness and understanding of data quality and the different data quality maturity dimensions which have an effect on data quality (DQ). In addition to the training session including experimenting and piloting the customized maturity model, the study includes observations done later via feedback about using the customized maturity model template and updating it further according to the comments. Additionally, literature review was conducted for taking into account the various aspects of data quality and data quality maturity models.

1.3 Case company

This thesis was done to an international ICT, network, and digital services company. The main driver for this thesis was the growing importance of data quality due to the continuously increasing amount of data and using it for internal and external purposes, e.g., reporting, automation, machine learning and artificial intelligence cases and decision making. Equally, or even more importantly, there is the need to improve the data quality in the source systems to ensure effective operations, customer and employee experiences, and ROI. In the case company, Gartner data quality maturity model *(Baskarada, 2009; Table 2.23: Bitterer, Gartner, 2007)* has been used earlier to share the maturity of the data quality and plan the actions for improving the data quality. Also, some other freely available data quality maturity checks have been done, for example Experian's Data Maturity Assessment (Experian, 2023) but the usage of data quality maturity assessments have not been included in the regular business processes and practices earlier.

1.4 Research limitations

New era of rapidly growing artificial intelligence and machine learning implementations have created even more need for transparency of the quality of data which is used as a basis of these implementations. Even while doing this thesis, more information about data quality related

implementations have been created, however not so much by academic publications but more by IT companies with commercial interests, and by specialized individuals. This information is publicly available, but it might not be that trustworthy academically. However, after carefully self-evaluating the validity of these, some of these were used as a reference in this study.

2 EFFECTS OF POOR DATA QUALITY FOR AN ORGANIZATION

There are studies about the effects of poor data quality already since 1990's, even from the 1980's (Zhu, Madnick, Lee, Wang., 2012). The article by Redman, published 1998, "*The impact of poor data quality on the typical enterprise*", is still cited a lot. It includes the typical impacts of poor data quality with three categories: Operational impacts, Tactical impacts, and Strategic impacts (Redman, 1998, Figure 2).

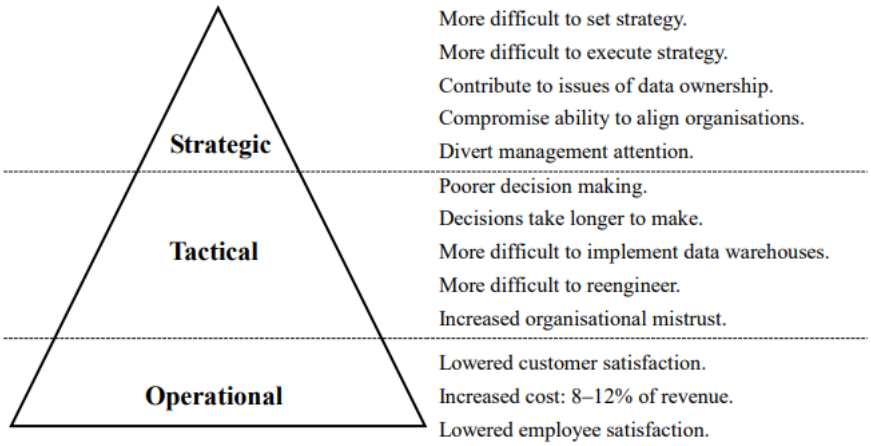


Figure 2. The Impact of Poor Information Quality (Baskarada, 2009, Redman 1998).

Now the amount of data has increased and is continuously increasing (Statista, 2023), impacts of poor data quality are increasing as well. In addition to the so-called traditional impacts for organizations, poor data has even wider effects nowadays because artificial intelligence and machine learning implementations are becoming a norm and the end-results of these are used even by a basic end-user. Whether these development efforts are useful from the end-user point of view, and worth the high investments for the organizations, depends on the quality of the data which is used in the development.

According to UK Government Data Quality Framework,

> "*If the stored data is unable to fulfil the requirements of its organization, then it is said to be of poor quality and it is deemed useless. All the time and effort an organization invests in capturing, storing, managing their data assets will be wasted if the data is not kept clean and error-free*." (UK Government Data Quality Framework, 2020)

In many cases, one data quality issue can affect to many below mentioned categories.

2.1 Financial impact and reduced productivity

According to research by Gartner, poor data quality costs businesses an average of $12.9 million per year (Gartner, 2021). Additionally, inefficiency and reduced productivity can occur if employees waste time managing data quality issues and correcting inaccurate or duplicate data instead of performing more productive tasks. This in turn creates extra costs for an organization, either as internal costs, or if done by external consultants, direct costs. Both costs can be either hidden or visible depending on the way of reporting these either as data quality costs, or if these costs are not reported as data quality costs, the costs of these tasks remain as hidden costs for the organization. When data quality issues affect customers, it leads to unnecessary contacts by them which also reduces the productivity and causes customer complaints. Redman has visualized the consequences of Department B correcting the data which it has received from Department A as a "*hidden data factory*" (Redman, Basecap, 2022, Figure 3). In this example, there are only two departments included but these hidden costs can be multiplied if there are more departments, and consequently, several "hidden data factories" in a company. Also, compliance violations to data privacy and data security regulations might take place and lead to penalties which can be very high. For example, GDPR regulators issued "*hundreds of fines to companies, including Google and Facebook, more than €114 million in the first 20 months of GDPR*" (GDPR, 2020).

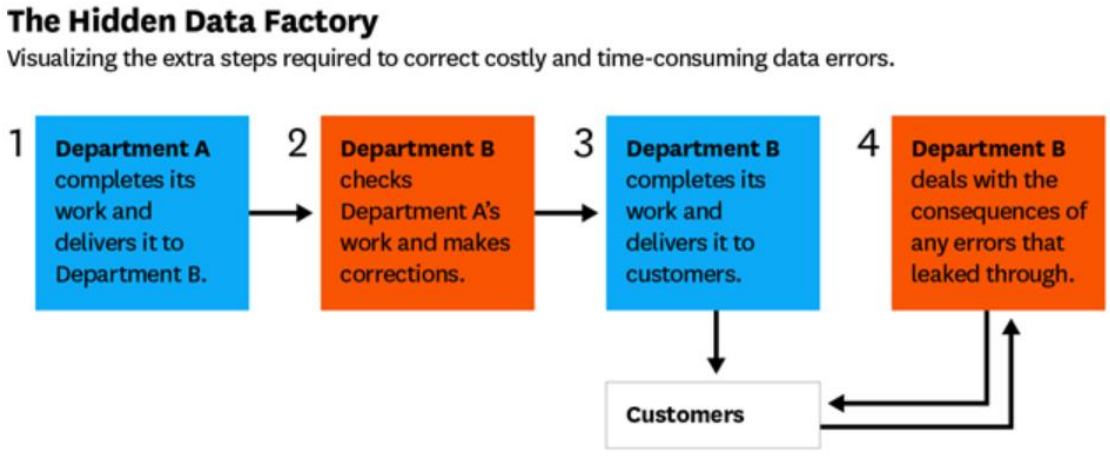


Figure 3. The Hidden Data Factory (Redman, Basecap, 2022)

2.2 Customer and employee experience

Customer experience might be compromised because inaccurate information being provided to customers, such as misspelled names, inappropriate product suggestions, undeliverable messages, duplicate communications, inaccurate transactions, and customer service histories. This can lead to leaving customers and poor Net Promoter Score (NPS) results. For the employees, it is frustrating to work with poor quality data which takes time from the productive work, not only in the operative work but also development work including poor quality data. For example, cleaning the data can take up to 80% of data scientists' time and "*it's the problem data scientists complain about most*" (Redman, 2018).

2.3 Poor decision making

Poor data quality causes inaccurate analytics which can lead to misguided decision making. Without transparency to the data and the status of its data quality, decisions can be called data-driven but these decisions might be poor if done with inaccurate data.

2.4 Reputation damage

Poor data quality can damage a business's reputation and lead to poor customer relations. For example, duplicate or inaccurate records could result in some customers being contacted multiple times, while others are missed out. It can also create mistrust among stakeholders. This could potentially lead to loss of business or have an effect when establishing new partnerships.

2.5 Missed business opportunities

Data analytics with inaccurate data or data which is not available when it would be required, can lead to missed opportunities. Nowadays it is even more important to have accurate real-time or near real-time data available for taking advantage of all the possible business opportunities.

2.6 Inaccurate data for automation and AI/ML cases

> "*Data powers almost all critical, customer-facing flows at Uber. Bad data quality impacts our ML models, leading to a bad user experience (incorrect fares, ETAs, products, etc.) and revenue loss."* (Uber, 2023)

In short, automating with poor quality data automates the creation of poor-quality data. For the AI/ML cases, data quality has to be even more accurate, including the historical data for developing predictive models in order to make reliable and well performing models.

> "*First, the data must be right: It must be correct, properly labeled, de-deduped, and so forth. But you must also have the right data — lots of unbiased data, over the entire range of inputs for which one aims to develop the predictive model."*
> (Redman, 2018)

Without good quality data available for AI/ML cases, the results of these implementations will be inaccurate and biased and can lead to poor decision making and wasted resources.

2.7 Delays in system migrations and/or poor data migrated to new system

There is very high risk for delays for the system migrations when having poor quality data in the legacy systems. It causes delays in the migration project and another probable cause is the poor data will be migrated to the new system as well thus causing extra problems there. It is like moving rubbish from an old apartment to a new one. The data in the legacy system should first be analyzed and corrected where needed. These actions must be considered already in the planning phase of the migration project because these actions probably have a considerable effect to the migration project schedule, resourcing, and costs. It is better to plan the project content and schedule accordingly than to postpone the go-live because of ignoring the poor data quality effects and costs already in the planning phase.

2.8 Impacts to sustainability and ESG reporting

Training AI/ML models take a lot of energy to run. According to MIT Technology Review, "*training just one AI model can emit more than 626,00 pounds of carbon dioxide equivalent – which is nearly*

five times the lifetime emissions of an average American car" (Hao, 2019; (Strubell, Ganesh, McCallum, 2019; Marr, Forbes, 2023). It is possible to estimate carbon footprints of AI models by using "*Machine Learning Emissions Calculator*" (Schmidt, Luccioni, Lacoste, Dandres, 2023). Consider a case where ML model includes poor quality data, and consequently, all those emissions are used in vain.

"*Just as you cannot create your financial report without accurate data on all the transactions, you cannot create your ESG report without accurate data on all the associated factors*" (Collibra, 2023). Reporting inaccurate carbon emissions and unreliable data on labour practices in ESG reporting, effect the ESG score. Via this, it effects strategic planning, investor decisions and corporate actions for ESG (Collibra, 2023). In addition to decreased sales and damaged brand reputation due to inaccurate ESG reporting, fines and penalties can also occur which also can have a substantial financial impact (U.S. Securities and Exchange Commission, 2022). Term called "*greenwashing*" means "*some companies are trying to lure investors in with false claims about their sustainability practices"*, (Forbes, 2022). Since greenwashing is a growing problem, and regulators are following this up even more carefully, it is now even a bigger risk for a company to include inaccurate data in their ESG reporting.

3 DATA QUALITY DEFINITION AND DIMENSIONS OF DATA QUALITY

ISO/IEC 25012 standard defines data quality as "*The degree to which data satisfies the requirements of its intended purpose*" (ISO 25012, 2022). However, there is no worldwide definition of data quality. According to Thomas C. Redman, "*data is considered to be of high quality if they are fit for their intended use in operations, decision making and planning*" (Redman, 2016). In addition, nowadays there are also increasing amount of compliance requirements for data so data quality should also address to this demand in order to avoid fines and penalties from the regulators.

Data quality dimension as a term is widely used to describe the measure of the quality of data. However, the key technical data quality dimensions are not standardized. Dimension is also used as a term in business intelligence where it "refers to a category for summarizing or viewing data." In their research "Dimensions of Data Quality (DDQ)" Van Nederpelt & Black have analysed the data quality dimensions and their definitions. As a result of this research, they have created a list of the preferred dimensions, and from these, they have selected most common DQ dimensions as Accuracy, Availability, Clarity, Completeness, Consistency, Currency, Punctuality, Timeliness, Traceability, Uniqueness, Validity (Van Nederpelt & Black, 2020).

3.1 Technical DQ dimensions

Accuracy "is the degree to which data represent real-world things, events, or an agreed-upon source" (Gawande, 2022). Accuracy is best to be validated at the source system, and/or measured and compared, against the correct information like reference data if available i.e. ISO country codes (ISO 3166, 2020), ISO currency codes (ISO 4217, 2015) or product codes by a company (ScienceDirect, 2021).

Completeness "is defined as the percentage of data populated vs. the possibility of 100% fulfilment" (Gawande, 2022). Gawande points out there can be 4 different types of data quality issues found in the completeness dimension:

1. **Missing record**, for example customer is totally missing from the customer database i.e. there is no information about a customer in the customer database although there should be.

2. **Null attribute**, for example there might be some mandatory information missing e.g. e-mail address or phone number (Figure 4).

Name	Email	Phone	Spend	Visit count	Reward points
Hank Williams	hank@msn.com	(564)342-1212	$210.03	2	47.14
Joe Panik		1415)321-7689	$37.45	1	35.00
David R Simcoke	devid@gmail.com		$59.13	2	30.00
R Kelly	rkelly@aol.com	(310)789-0000	$24.64	2	27.28
Bruce Bocily	bbochy@stgiants.com	(415)456-7890	$0.00	0	26.79
Buster Posey	buster@orgiants.com		$261.20	12	26.06
Klay Thompson	splashbro@Pwarriomerun	(510)543-2345	$0.00	0	25.24
Steve Kerr			$268.53	13	24.25
Ayeesha Curry	acurry@gmail.com	(510)426-7457	$407.52	8	19.70
Tim Lincecum	timmy@sfgiants.com	(415)453-2345	$3.272.99	126	19.23
P Diddy	diddy@outlook.com	(510)765-6789	$0.00	0	17.95

Figure 4. Examples of missing attributes like e-mail addresses and phone numbers. (Gawande, 2022)

3. **Missing reference** data means for example, all the required reference values are not available to be selected when it is mandatory to select one of the values from the list. For example there are only values between 1-10 but the user would need to select 15 and it is not available to be selected.
4. **Data truncation**, refers to a situation where for example when only part of the data is uploaded to database because the target attribute is not large enough compared to the source data attribute (Figure 5).

MGR	HIREDATE	JOB
7902	17-DEC-80	CLERK
7698	20-FEB-81	SALES
7698	22-FEB-81	SALES
7839	02-APR-81	MANAG
7698	28-SEP-11	SALES
7839	01-MAY-81	MANAG
7839	09-JUN-81	MANAG
7566	19-APR-87	ANALY
	17-NOU-81	PRESI
7698	08-SEP-81	SALES
7788	23-MAY-87	CLERK
7698	03-DEC-81	CLERK
7566	03-DEC-81	ANALY
7782	23-JAN-82	CLERK

JOB
CLERK
SALESMAN
SALESMAN
MANAGER
SALESMAN
MANAGER
MANAGER
ANALYST
PRESIDENT
SALESMAN
CLERK
CLERK
ANALYST
CLERK

Figure 5. Examples of truncated data values in Job -column. (Gawande, 2022)

Consistency, according to the definition by Gawande refers to "how close your data aligns or is in uniformity with another dataset or a reference dataset." (Gawande, 2022). His examples include:

1. **Record inconsistency**: Record which exists in the source system, cannot be found in the target system.
2. **Attribute inconsistency**: Records exist in both source and target databases/systems, but their attributes do not match with each other.
3. **Data inconsistency over time**: Remarkable change in one attribute value or data volume compared to the expected minor variations. For example, stock price of one company suddenly increases by 10 times, or the average of new customers for a company per day is about 500 but suddenly there seems to be thousands of new customers or the customer count suddenly drops to zero
4. **Reference data inconsistency**: Reference data is stored and used inconsistently in different datasets and systems. For example, for the country codes, many different values are found even though reference data for using the country codes has been provided but it is not validated in the systems and datasets.

Timeliness according to Gawande, "is the time lag between actual event time vs. the event captured in a system to make it available for use" (Gawande, 2022). Data quality of timeliness depends

on user expectation for example for the financial results the accuracy of the figures might be of good quality but if these figures are not available on time for the report, the data quality of the timeliness is poor.

SLA	Table Load Time
08:00 am	07:59 am
10:00 am	09:59 am
11:00 am	11:01 am

← Missed the SLA

Figure 6. Example about timeliness data quality issue. (DataCamp, 2023).

Relevance, in short, means the data is applicable to what it is being used, for example for solving a business question or a problem. There has to be information about what the data is suitable to be used for.

Validity means the data is accurate and the degree how close the data value is to predetermined values or calculation (Gawande, 2022). It can also mean the data values are in the ranges which have been provided in advance, like below 100 or between 100 to 200. Examples from DataCamp about validity include the customer's birthdate must be a date in the past, customer's account type must be either "*Loan*" or "*Deposit*" (Figure 7).

CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

Figure 7. Examples about validity data quality issues (Datacamp, 2023).

Reliability means the data represents what it claims to represent also over time and you can trust it to be valid also in the future.

Integrity is, as defined by Gawande, "the degree to which a defined relational constraint is implemented between two data sets". Data integrity issues can be found in one system or between several systems (Gawande, 2022).

Precision, according to Gawande, is defined like "The degree to which the data has been rounded or aggregated" (Gawande, 2022). There can be for example numerical precision errors like rounding of number with too few digits. For example when there using a two-digit precision for GPS coordinates the error can be one kilometre compared to using five-digit precision where error is only one meter. For the time precision, the simple example is when using only a date to record a purchase when also the time would be required in order to get more precise information about the time of purchase. (Gawande, 2022).

Format means the "data values of the same attributes must be represented in a uniform format" (Gawande, 2022). Gawande also defines this as format conformity and gives the date format as an example, for example the same date in a format as 'YYYY/MM/DD' and 'DD-MM-YY'. Gawande also mentions data type as another kind of conformity issue. This could mean a numeric value is expected but alpha numeric has been given. (Gawande, 2022).

Duplication (Uniqueness) means there are no duplicate records. Example of duplicate record is e.g. the same person is included in the customer database with different names or the same person is included in the customer database twice or multiple times.

3.2 Governance DQ dimensions

Governance type of data quality dimensions is not typically found in the data quality maturity assessment but these are more likely to be included in the data management or data governance related maturity models. However, governance is the foundation for improving data quality so in addition to the technical data quality dimensions, governance related data quality dimensions were also included in the customized data quality maturity measurement. Below are the definitions of the data governance related dimensions included in the customized maturity assessment.

Data quality governance involves processes for data quality improvement like a defined life cycle for recognizing, collecting, prioritising, and controlling data quality issues.

Data quality standards mean there are standards defined for data quality. Experian defines data quality standard as "a documented agreement on the representation, format and definition for common data" (Experian, 2023). It ensures the data entry values are unified. Data can then also be validated according to these standards.

Data quality metrics are the measurements of the quality of data. Measurements are done according to the business rules. Threshold limits can be decided for the data quality metric so when a certain threshold has not been reached, data quality issue will automatically be reported.

Data Quality issue reporting and mitigation is about reacting, prioritizing, and handling data quality issues. Data quality issues can be reported manually or automatically via data quality monitoring implementations. Data quality issue reporting gives visibility to the data quality issues and via this, also the root causes of the data quality issues can be analysed and corrected if possible. DQ issue corrections should be prioritised according to business value so for this purpose, there should be a way to prioritise them. A simple version can be created for example in MS Excel by using variables like the approximate financial benefit of correcting the issue and approximate costs of the correction. In the data catalogue tool IntoZetta, there is possibility to include and show the financial costs of errors for the business processes in the dashboard (Intozetta, 2023). With the latest tools using ML and AI, it could also be possible to do the corrections automatically for these technical DQ issues.

Data quality processes exist for taking care of the data quality proactively. Processes and practices have been established for taking data quality into account in all parts of the processes in the organization. Data quality issues are followed-up, even predicted and corrected, so these do not come as a surprise but are managed in a controlled way.

General data management is a way to collect, organize and store the data of the organization. The goal is to help organizations to use the available data to make decisions and take actions for the benefit of the organization. For this reason, data assets should also be documented, preferably in a data catalogue.

3.3 Culture & competencies DQ dimensions

Culture and competencies are very important factors for improving data quality. So, in addition to Governance, these were also added to the first version of the customized data quality maturity assessment as DQ dimensions.

Collaboration: Silo / Group work as a DQ dimension, mean that especially in a bigger company, data quality cannot be improved in silos. It is quite typical the data producers do not even know the requirements for the data quality of the data consumers and data quality cannot be improved either. There has to be collaboration and information sharing across the organization about the data quality. That way also the importance about the data quality management gets more visibility and priority. This may hopefully also lead to the management understanding about the need to include data, and data quality, related competence development in the business strategies and KPIs.

Self-service capabilities mean there are possibilities for the business users to interact with the data without the need to order and get the data from the technical personnel. Especially in bigger companies, it is often a problem the data consumers do not have access to the data on their own but need some technical persons to provide them data for example via SQL scripts, and often without any documentation about these scripts. In these cases where the documentation is missing, there is no visibility about what kind of rules have been used to fetch the data. Nowadays, whenever needed, data consumers should have access to the data themselves, and fetch the data they need without technical help. This is the way to improve the data literacy competencies and data quality from the business point of view in the most efficient way.

4 MEASURING AND ANALYSING TECHNICAL DATA QUALITY

Measuring data quality creates visibility to the status of data quality so the data quality is no longer a black box. Measuring data quality has not yet been standardised but it is done according to the needs and resources of each organization. In addition, artificial intelligence and machine learning have created a need for new measurements to be checked and monitored. For example, for the machine learning, both the quality of historical data quality and the quality of the new data effect on the machine learning models (Redman, 2018).

In his article, Redman did two statements of measuring data quality like the expected field error rates of about 1-5% meaning the percentage of the fields with an error from the number of total fields. Additionally, he states, if the measurements are not done, the organization probably has other serious data quality problems as well. (Redman, 1998)

4.1 Data quality measurement and monitoring tools

The goal for measuring the data quality should be to have ongoing measurements instead of one-time activities (Sebastian-Coleman, L., 2013). Even better target for measuring and monitoring data should be to get it automated so there will be automated notifications available when the decided data quality threshold has not been reached. In this case the ultimate target would be to also get the corrections done automatically whenever possible. This will be the future of the data quality management i.e. by taking advantage of the augmented data quality solutions which use artificial intelligence and machine learning (Forbes, 2022; Gartner, 2022). With the current speed and volume of data, it might even be impossible to do the required measurements without automation,

For automating data quality measurements and monitoring, there are several tools available. Ehrlinger and Wolfram conducted a survey about the data quality measurement and monitoring tools (Ehrlinger and Wolfram, 2022). In their survey, they focused on the measurement and automated data quality monitoring capabilities of these tools. They were able to find altogether 667 data quality tools and defined the evaluation criteria to these tools which included the following functionalities: data profiling, data quality metrics creation and automated data quality monitoring. By using this criterion, they were able to select 8 commercial and 5 open-source tools for more detailed analysis

(Ehrlinger, Wolfram, 2022). However, it depends on each organization's needs and resources which could be considered the best fit for them.

4.1.1 Magic quadrant for data quality solutions by Gartner

Gartner has done research on the commercial data quality solutions and published "*Magic Quadrant for Data Quality Solutions*" as a result (Figure 8). In Gartner's research, the focus is on these solutions which can serve not only current but also the future needs of the data quality use. According to Gartner, evaluation and selection of the data quality solutions is no more specialized but requires collaboration with business and these who have plans to use these solutions for their various use cases (Gartner, 2022). Gartner states the data quality solutions are transitioning to augmented data quality solutions which they define "*as set of capabilities for enhanced data quality experience aimed at improving insight discovery, next-best-action suggestions, and automation by leveraging AI/ML features, graph analysis and metadata analytics*". (Gartner, 2023).

Figure 8. Magic Quadrant for Data Quality Solutions by Gartner (Gartner, 2022).

4.2 Profiling data

Data profiling is a statistical analysis and way to get more information about dataset and the status of data quality of a dataset. Via data profiling it is possible to get a column specific information about e.g., the number of missing i.e. null values, number of distinct values, data types, formats, minimum and maximum lengths and values et cetera.

Most data quality tools have at least a basic level of data profiling capability. In addition to the data quality tools, there are also Python libraries available for data profiling, for example ydata-profiling

(alexbarros, 2023), lux (RenChu Wang, dorisjlee, 2023), DataProfiler (taylorfturner, 2023) and Great Expectations (Great Expectations, 2023).

Other tools mentioned in Gartner magic quadrant (Gartner, 2022) offer similar tools, with slightly varying options. Tools such as IBM SPSS Modeler have had graphical user interface operated data profiling tools for over a decade (IBM, 2023).

Example from Experian Aperture Data Studio about profiling with Core statistics (Figure 9). In Experian example the core statistics in their profiling feature are: Uniqueness, Completeness, Row Count, Has Nulls, Dominant Datatype, Format Count, Shortest Length, Longest Length, Rare Values, Frequent Values, Long Values and Missing Values (Experian, 2023).

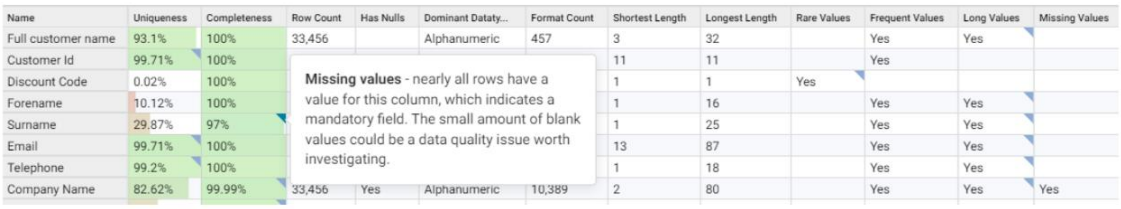


Figure 9. Example of Experian Aperture Data Quality tool data profiling functionality (Experian, 2023).

In their data quality tool, they have altogether 57 profiling statics, and due to their column tagging feature, profiling feature also includes the data tags and sensitive data tags. They have also enhanced their profiling features with the notes for the users about the possible data quality issues and automated validation rules based on these notes (Experian, 2023). The Figure 10 illustrates the amount of details which could potentially cause data quality issues for a dataset. By using the data profiling functionalities, these can be visualized as a summary and these details can be checked according to the intended purpose of the dataset.

Attribute	Description
Unique Count	The number of unique values in the column (same as the count of grouped values. Nulls are counted as a value).
Minimum	The minimum value in the column (first alphabetically, earliest date, lowest numeric value).
Maximum	The maximum value in the column (last alphabetically, latest date, highest numeric value).
Overall Datatype	The Datatype that would be required to store all values. If a column has both Numeric and Date datatypes, the overall type would be Alphanumeric.
Sum	The sum of all numeric values in this column.
Standard Deviation	The standard deviation of all numeric values in this Column. Non-numeric values are treated as zero.
Average	The average of all numeric values in this column (sum of numbers / number count).
Precision	The overall numeric precision for this column, which is the largest number of digits in a number including digits on both sides of the decimal point.
Scale	The overall numeric scale for this column which is the most number of digits to the right of the decimal point.
Zero Count	The number of rows in the Table containing a value in this column that is zero
Negative Values	The count of numeric values in this column that are less than zero
Least Common Value	The overall least frequently occurring value in the column.
Least Common Count	The count of the number of time the least common value occurs in the column.
Most Common Value	The overall most frequently occurring value in the column.
Most Common Count	The count of the number of time the most common value occurs in the column.
Least Common Format	The least frequently occurring format pattern in this column.
Least Common Format Count	The count of the number of times the least frequently occurring format pattern occurs in this column.
Most Common Format	The most frequently occurring format pattern in this column.
Most Common Format Count	The count of the number of times the most frequently occurring format pattern occurs in this column.
Average Length	The average lengths of all values in the column.
Length Deviation	The standard deviation of the lengths of all values in the column.
Frequency Deviation	The standard deviation of the frequency of occurrence of each value across the set of values in the column.
Format Frequency Deviation	The standard deviation of the frequency of occurrence of each format pattern across the set of format patterns in the column.
Alphanumeric Uniqueness	The uniqueness of the alphanumeric values in the column (unique alphanumeric values as a percentage of all alphanumeric values).
Alphanumeric Unique Count	The number of unique alphanumeric values in the column.
Alphanumeric Completeness	The completeness of the alphanumeric values column (values in the column that are alphanumeric as a percentage of the total row count).
Alphanumeric Count	The count of alphanumeric values in the column.
Alphanumeric Minimum	The first alphanumeric value in the column alphabetically.
Alphanumeric Maximum	The last alphanumeric value in the column alphabetically.
Number Uniqueness	The uniqueness of the numeric values in the column (unique numeric values in the column as a percentage of all numeric values).
Number Unique Count	The number of unique numeric values in the column.
Number Completeness	The completeness of the numeric values column (values in the column that are numeric as a percentage of the total row count).
Number Count	The count of numeric values in the column.
Number Minimum	The lowest numeric value in the column.
Number Maximum	The highest numeric value in the column.
Check Sum	The checksum for all the values in the column.
Date Uniqueness	The uniqueness of the date values in the column (unique date values in the column as a percentage of all date values).
Date Unique Count	The number of unique date values in the column.
Date Completeness	The completeness of the date values column (values in the column that are date as a percentage of the total row count).
Date Count	The count of date values in the column.
Date Minimum	The earliest date value in the column.
Date Maximum	The latest date value in the column.
Nullity	The percentage of rows in the table where the value in this column is null.
Null Count	The number of rows in the table containing a value in this column that is null.
Key Check	Whether the data in the column denotes this is a perfect key or is a key that is broken.
Rare Formats	Whether there are unexpected infrequent formats in the column.
Short Values	Whether there are abnormally short values in the column.
Low Amounts	Whether there are abnormally low numeric values in the column.
High Amounts	Whether there are abnormally high numeric values in the column.
Sequence	Whether the values in the column are sequential numbers.
Average Frequency	The average frequency of occurrence of a value across the set of values in the column.
Average Format Frequency	The average of frequency of occurrence of a format pattern across the set of values in the column.
Sum Squared	The sum of the squares of all numeric values in the column
Length Sum Squared	The sum of the square of lengths of all values in the column.
Length Sum	The sum of the lengths of all values in the column.
Sum Squared Of Frequency	The sum of the square of frequency of occurrence of each value in the column.
Sum Squared Of Format Frequency	The sum of the square of frequency of occurrence of each format pattern of values in the column.
Data Tags	Data tags assigned to the column.
Sensitive Data Tags	Data tags with sensitive flag assigned.

Figure 10. Experian Aperture Data Quality tool data profiling attributes and their descriptions (Experian, 2023).

According to Ehrlinger and Wolfram, data profiling is an essential task before creating data quality metrics in order to get insight of the details of a dataset in question (Ehrlinger, Wolfram, 2022). Getting more insights of the datasets, and their quality, is equally important before data analysis, data migrations and data warehousing so data profiling should be considered as a default practice for these purposes as well.

4.3 Creating metrics

Data quality metrics have to be created according to the business benefits. Obviously, there is no point to waste time and resources in checking and correcting fields which have no business value. The simplest way for creating data quality metrics would be to create metric(s) for checking data against for example, reference or master data values in business-critical data assets and their fields. Typically, the first metrics which are created are checking for the null values and incorrect formats, and after this, develop the metrics incrementally further by adding more checks and validations including business related rules.

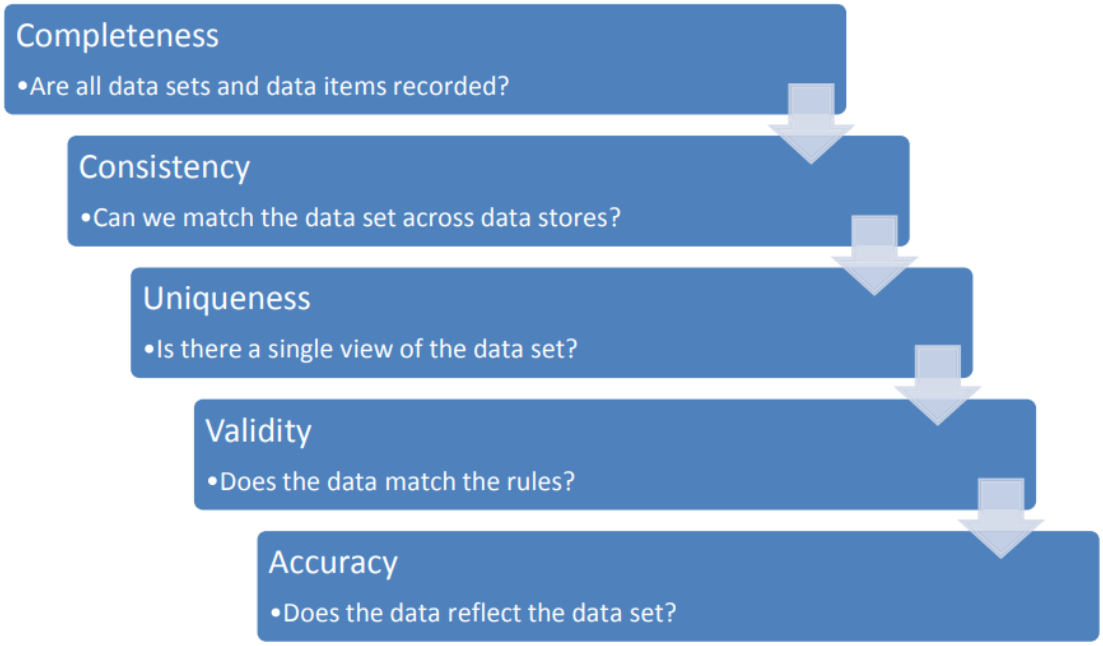


Figure 11. Example of applying data quality dimensions to a dataset by DAMA UK Working group (DAMA UK, 2013).

4.4 Monitoring and detecting data issues

> *"Still, many data issues are manually detected by users weeks or even months after they start. Data regressions are hard to catch because the most impactful ones are generally silent. They do not impact metrics and ML models in an obvious way until someone notices something is off, which finally unearths the data issue. But by that time, bad decisions are already made, and ML models have already underperformed.*

This makes it critical to monitor data quality thoroughly so that issues are caught proactively." (Uber, 2023)

Once data quality metrics have been created according to the DQ rules based on business benefits, the metrics should be continuously and automatically monitored in order to catch the data quality issues quickly. Data quality issues should be automatically notified including the outliers, anomalies, patterns and drifts. Additionally, target would be also to provide monitoring dashboard, log files or audit trail for compliance requirements. This kind of dashboard should also provide visibility to interactive analytical workflow and visual output of statistical analysis to help business and IT users identify, understand, and monitor data quality issues and discover patterns and trends over time.

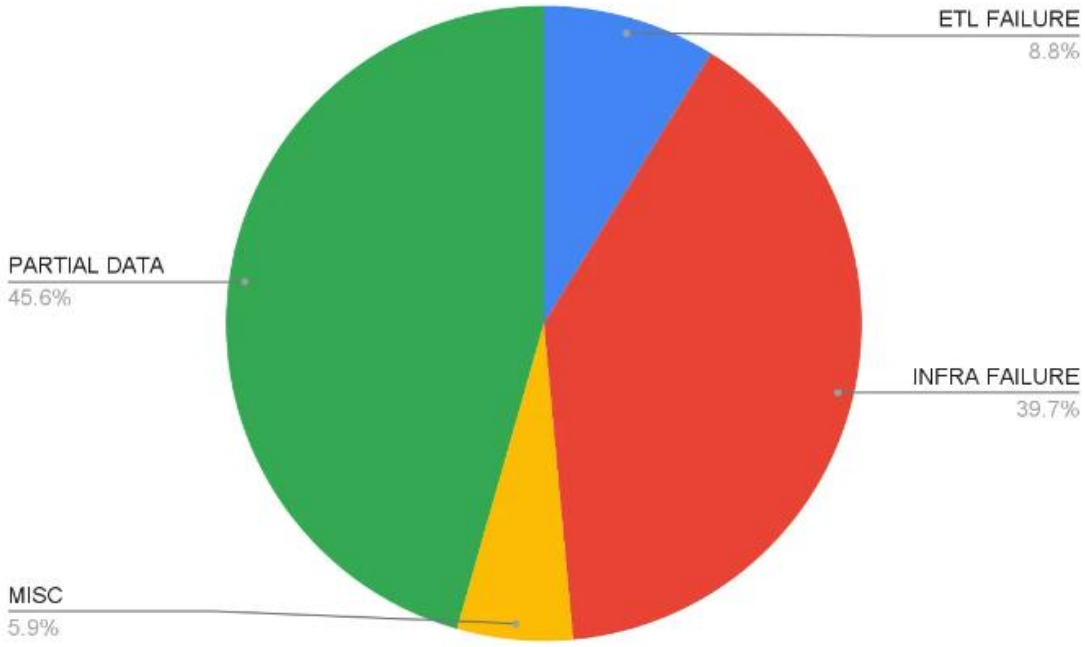


Figure 12. Example from Uber of the data incidents by categories in 2022 (Uber, 2023).

4.5 Augmented data quality management

"*So much data is being generated and consolidated so quickly that it's become impossible to use traditional methods to manage data quality successfully*." (Forbes, 2022)

Augmented data duality management means automating data quality processes and practices by taking advantage of advanced algorithms, machine learning and artificial intelligence. It enables to

also correct data automatically, learn from that and improve according to the learnings. Augmented data quality management is a must have in near future, and "*machine learning and automation can reduce manual data management tasks by 45 percent*." (Deloitte, 2022). Augmented data quality can improve data quality productivity but it is not possible to fully automate it. However, it is not a must to have 100% data quality so it can be so that the machine learning model could achieve the accepted level of data quality by being able to detect enough data quality problems and correct these automatically. Self-learning models could get even better results but, in any case, data stewards are needed to validate or review and accept the corrections. Data stewards understand the business needs of the data and because of this, the corrections cannot be fully automated. Even though ML and AI applications can help improve the data quality, people like business data stewards, are still needed to be involved in the data quality activities.

4.6 Root causes of data quality issues

Redman has stated "*To Improve Data Quality, Start at the Source*" (Redman, 2020).

> "*Rather than fixing data quality by finding and correcting errors, managers and teams must adopt a new mentality — one that focuses on creating data correctly the first time to ensure quality throughout the process. This new approach — and the changes needed to make it happen — must be step one for any leader that is serious about cultivating a data-driven mindset across the company, implementing data science, monetizing its data, or even simply striving to become more efficient. It requires seeing yourself and the role you play in data in a new way, all the while identifying and ruthlessly attacking the root causes of errors, making them disappear once and for all*." (Redman, 2020).

Techniques for identifying root causes of the data quality issues are the same as for any root cause analysis, like Fishbone and 5-Whys as presented with good example cases by Southekal (Southekal, 2022). This example fishbone, in fact, includes quite many of the root causes of data quality issues like lack of data standards, poor data integration and using free text fields. To find out the root cause of each DQ issue, first select a data quality issue for which there is a need to find a solution, and then use for example fishbone diagram (Figure 13) and/or 5-whys method to first find out a root cause(s) and then plan how to fix these.

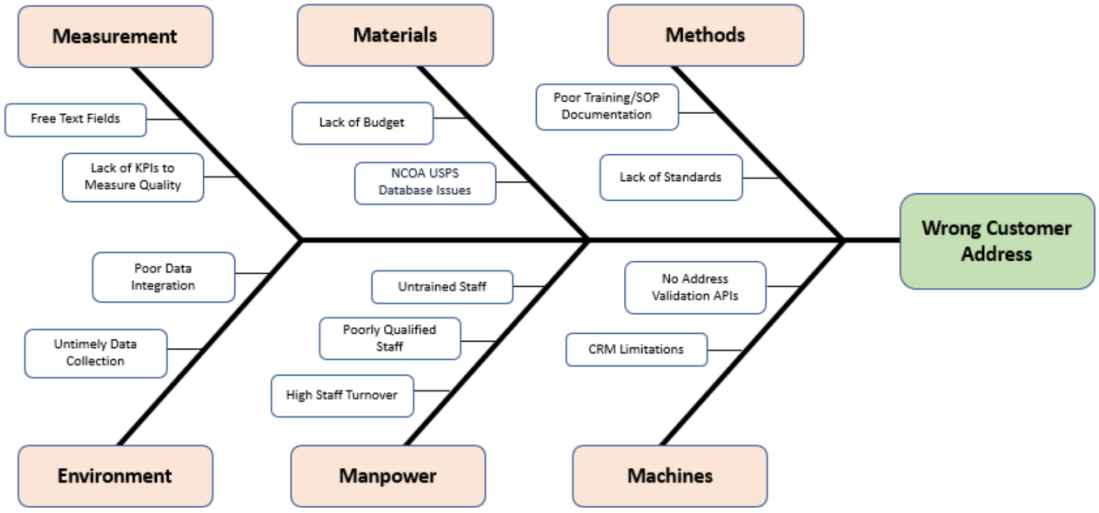


Figure 13. Fishbone diagram example by Southekal for identifying a root cause for wrong customer addresses (Southekal, 2022).

For drilling down to the individual data quality issues, after the fishbone diagram, Southekal suggests to use 5-whys method i.e. five consecutive questions starting with why. (Southekal, 2022).

As an example of 5-whys, he used the following question set:

"*1. Why didn't the Data Pipeline routine get deployed on time?*

– Because the development could not be completed on time.

2. Why were the development not completed on time?

– Because the testing the application took a lot of time.

3. Why did testing the application took a lot of time?

– Because there was no quality data available to test.

4. Why was data quality poor?

– Because data was manually entered by poorly trained users.

5. Why were the users not trained?

– Because we do not have a data literacy program in the company."

(Southekal, 2022).

According to Redman two most frequent root causes involve:

1. Data Producers, i.e. those who create data, do not know others have requirements for their data

2. Data Consumers, i.e. those who are victimized by poor data, reflexively act to fix bad data (Redman, 2022)

The validation and/or monitoring of the data at each source system should also consider the further usage of this data to prevent the consequences of poor data quality in downstream. Should these data quality issues arise for the data consumers, they should be reported so that the root causes for these could be corrected.

Another example of root cause is from Uber where they had an app experiment which started to log fares differently (Uber, 2023).

Additionally, root cause example from Statistics Finland:

> "*The rise in prices was taken into account incorrectly twice in the price index of electricity. The rise in prices was for the first time introduced in the index in January 2022, when the rise in prices was visible as an increase in the obligation to deliver prices, and for the second time the corresponding price rise visible to consumers was included in the index in November 2022*." (Statistics Finland, 2023)

From these examples, it can be learned there is a high risk for data quality problems when changes in implementations occur. All the changes must be carefully planned and tested, in the similar manner as in the software development process. Continuous data quality monitoring should also take place so any changes in the data can quickly be noted and considered in the implementations.

5 OVERVIEW OF DATA QUALITY MATURITY TOOLS

Maturity models can be used as a tool to benchmark the current state of an organization and guide it how to improve and get to the next level of maturity. Typically, data governance, data management and data quality maturity models are presented in quite high-level stages which do not offer sufficiently differentiation between the maturity stages so it is challenging to use these for checking and sharing the information about the improvements in the maturity. When such high-level maturity model is used, it can take years to reach the next maturity level and make the progress visible in this kind of model.

Several data management and data governance maturity models have been published and marketed by consulting companies. Most of these need to be purchased. There is research done via literature review which includes an overview of 22 maturity models for data management (Belghith; Zitoun; Skhiri dit Gabouje; Ferjaoui 2021). For example, TDWI is included in their study and TDWI has published a new data management maturity model assessment in the beginning of 2023 and it can be used free of charge by giving the contact details (TDWI, 2023). However, this maturity model includes high level questions, and mostly other than data quality maturity questions so it does not include data quality maturity score separately but in addition to Overall score, it includes scores for Organization, Resources, Architecture, Data lifecycle and Governance (TDWI, 2023).

Comparing different maturity models is challenging, or even impossible. In her article, data management practitioner Irina Steenbeek compared between DAMA-DMBOK2 and DCAM® 2.2 models and found there are the same levels included, i.e., Level 0 to Level 5 but these levels have so different descriptions they cannot be compared (Figure 14).

Level	Level name		Level description	
	DAMA-DMBOK2	DCAM® 2.2	DAMA-DMBOK2	DCAM® 2.2
Level 0	No capability	Non initiated	No organized data management practices or formal enterprise processes for managing data	Ad-hoc data management (performed by heroes)
Level 1	Initial/Ad Hoc	Conceptual	• Little or no governance • Limited tool set • Roles defined within silos • Controls applied inconsistently • Data quality issues not addresses	Initial planning activities (white board sessions)
Level 2	Repeatable	Developmental	• Emerging governance • Introduction of a consistent tool set • Some roles and processes defined • Growing awareness of impact of data quality issues	Engagement underway (stakeholders being recruited and initial discussions about roles, responsibilities, standards and processes)
Level 3	Defined	Defined	• Data viewed as an organizational enabler • Scalable processes and tools • Reduction in manual processes • Process outcomes are more predictable	Data management capabilities established and verified by stakeholders (roles and responsibilities structured, policy and standards implemented, glossaries and identifiers established, sustainable funding)
Level 4	Managed	Achieved	• Centralized planning and governance • Management of risks related to data • Data Management performance metrics • Measurable improvements in data quality	Data management capabilities adopted and compliance enforced (sanctioned by executive management, activity coordinated, adherence audited, strategic funding)
Level 5	Optimization	Enhanced	• Highly predictable processes • Reduced risk • Well understood metrics to manage data quality and process quality	Data management capabilities fully integrated into operations (continuous improvement)

Figure 14. Comparison between DAMA-DMBOK2 and DCAM® 2.2 maturity models (*Steenbeek, 2021).*

5.1 Gartner's data quality maturity model

Gartner has published Data Quality Maturity Model in 2007 and it is available for Gartner clients, and it can also be purchased on their website (Gartner, 2007).

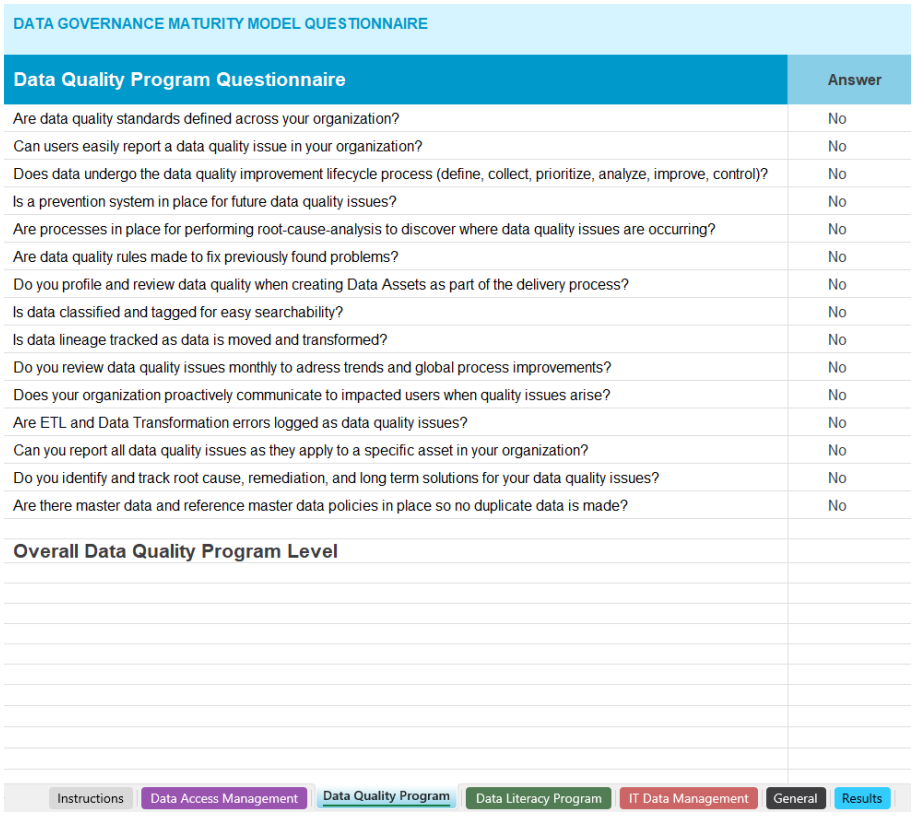
Sasa Baskarada has included Gartner's Data Quality Maturity Model in his research about Information Quality Management Capability Maturity Model (Figure 15). These levels are defined in high level and by following this, it would be challenging to set targets and follow-up the progress by using this and get company-wide visibility to the data quality maturity. This could maybe help to guide the data quality improvement actions in quite a small company but for the bigger ones, it would need to be more detailed and/or followed-up per business unit. For this reason, by using this one, it could take years until the next level has been reached. Additionally, there is nothing mentioned about the business benefits versus data quality actions. Data quality improvement actions must be prioritised in such a way the business benefits will be gained out of these, and the sooner, the better.

Level 1: Aware Very little understanding of DQ concepts. Any problems are largely ignored. No formal DQ initiatives. No incentives to improve DQ. General ignorance and belief that all information is correct by default. No DQ responsibilities or accountabilities.
Level 2: Reactive Data formats, mandatory fields, and value validations are being enforced. Some limited manual batch cleansing performed. Users are waiting for problems to occur, instead of taking proactive steps. DQ problems are still perceived to be solely the IT department's responsibility.
Level 3: Proactive DQ gradually becomes part of the IT charter. DQ tools (e.g. profiling & cleansing) are in use. DQ is considered "good enough" for most decision-making. Major issues are documented, but not completely rectified. DQ guidelines and data ownership are emerging.
Level 4: Managed DQ is a prime concern. Commercial DQ software implemented more widely. Regular DQ assessments and impact analyses performed. DQ functionality is introduced beyond business intelligence and data warehousing programs. Multiple data stewardship roles are established. Metrics-based DQ dashboards in use.
Level 5: Optimising Rigorous processes are in place to keep DQ as high as possible, through ongoing housekeeping exercises, continuous monitoring of quality levels, and by attaching quality metrics to the compensation plans of data stewards and other employees. DQ becomes an ongoing strategic initiative. Subjective DQ assessments (e.g. believability, relevance and trust factors). Data is enriched in real time by third-party providers.

Figure 15. Table 2.23: Gartner's Data Quality Maturity Model, Bitterer, Gartner, 2007 (Baskarada, 2009).

5.2 OvalEdge data governance maturity model

OvalEdge offers data governance maturity model questionnaire which can be used free of charge by giving the contact details. It includes questionnaire about data quality program (OvalEdge, 2023, Figure 16). Additionally, it includes questionnaires about data access management, data literacy program, IT data management and general data governance questions (OvalEdge, 2023). It is a concrete and easy way to assess Data Quality and Data Governance maturity without having to purchase this kind of tool. Compared to e.g. TDWI, DAMA-DMBOK2 and Gartner Data Quality Maturity models, it also includes more in-depth questions about the matters effecting to data quality and data governance maturity so it could be used to guide the organization towards these concrete actions which are needed to improve data quality and DQ maturity. By using this tool, it would be possible to follow-up and get visibility to the status of data quality improvement actions and it would not take years to see the progress of data quality improvement actions via using this tool. However, it is still in quite a high level and like with the Gartner DQ Maturity Model (Baskarada, 2009; Table 2.23: Bitterer, Gartner, 2007), it would probably be more useful in quite small companies. Additionally, this one also lacks the business benefits perspective of the DQ improvement actions.



DATA GOVERNANCE MATURITY MODEL QUESTIONNAIRE

Data Quality Program Questionnaire	Answer
Are data quality standards defined across your organization?	No
Can users easily report a data quality issue in your organization?	No
Does data undergo the data quality improvement lifecycle process (define, collect, prioritize, analyze, improve, control)?	No
Is a prevention system in place for future data quality issues?	No
Are processes in place for performing root-cause-analysis to discover where data quality issues are occurring?	No
Are data quality rules made to fix previously found problems?	No
Do you profile and review data quality when creating Data Assets as part of the delivery process?	No
Is data classified and tagged for easy searchability?	No
Is data lineage tracked as data is moved and transformed?	No
Do you review data quality issues monthly to adress trends and global process improvements?	No
Does your organization proactively communicate to impacted users when quality issues arise?	No
Are ETL and Data Transformation errors logged as data quality issues?	No
Can you report all data quality issues as they apply to a specific asset in your organization?	No
Do you identify and track root cause, remediation, and long term solutions for your data quality issues?	No
Are there master data and reference master data policies in place so no duplicate data is made?	No
Overall Data Quality Program Level	

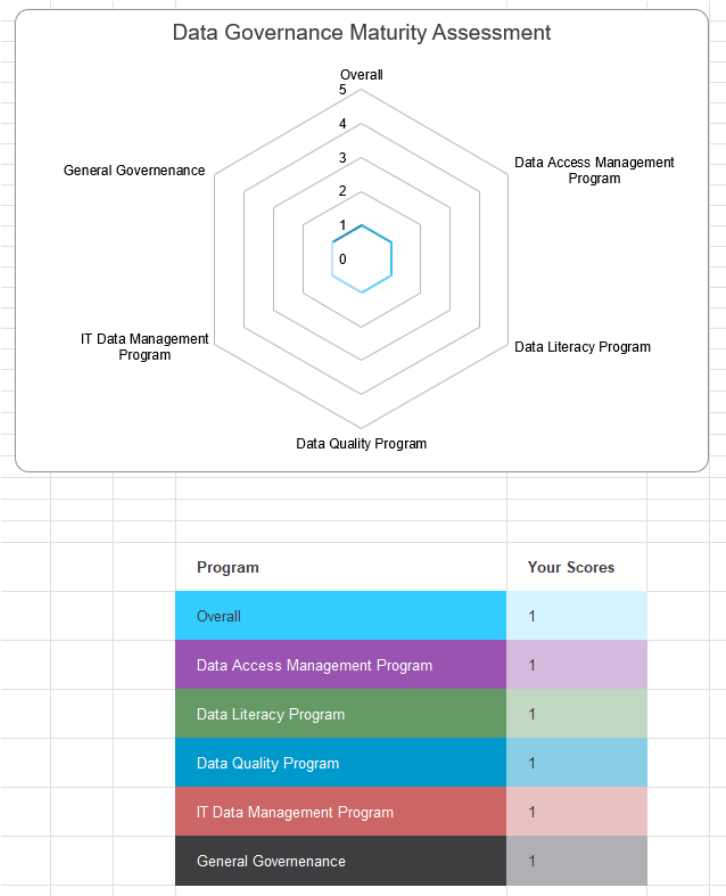
Figure 16. Data Quality Program Questionnaire sheet of OvalEdge Data Governance Maturity model which includes 15 general questions about data quality (OvalEdge, 2023)

Questions in the OvalEdge data governance maturity model are to be answered in levels 1-5 with the following values (OvalEdge, 2023):

Level	**Answer option to maturity model questions**
Level 1: No	"*Issues are dealt with as they appear*"
***Level 2:* Beginning**	"*The importance is becoming apparent and commonly accepted. Efforts are beginning as the organization determines what is needed*"
Level 3: In Progress	"*Policies and documentation are being created to implement a solution. Actions are being taken such as appointing people to positions or installing a tool.*"
Level 4: Yes	*"The organization is now enforcing policies and procedures for data governance. There is an implemented solution being monitored for success.*"

Level 5: Absolutely	"*The implemented solution is working and only needs a small improvements for optimization. Workflows are re-designed to reduce redundancy*."

After completing the OvalEdge Data Quality Program questionnaire (OvalEdge, 2023), the results are visualized in the "Results" sheet (Figure 17). In order to get the overall results for Data Governance Maturity, also the other related questionnaires should be completed.



Program	Your Scores
Overall	1
Data Access Management Program	1
Data Literacy Program	1
Data Quality Program	1
IT Data Management Program	1
General Governenance	1

Figure 17. Results sheet of OvalEdge Data Governance Maturity questionnaire (OvalEdge, 2023).

6 CUSTOM DATA QUALITY MATURITY MEASUREMENT ASSESSMENT

Data quality maturity model must suit organization's needs, goals, and context so it is also possible to customize an existing model to fit to these specific requirements of an organization. Customized data quality maturity measurement tool was used in this study. The first draft version of this was based on existing literature like the book “Data Quality Assessment” (Maydanchik, 2012), “Bad Data Handbook” (McCallum, 2012) and the Gartner DQ Maturity Model (Baskarada, 2009; Table 2.23: Bitterer, Gartner, 2007). It was created in MS Excel as a heatmap by using four level scoring. Technical parts of the maturity measurement had initially been created for measuring the data quality maturity of handling the batch data and data fetched via integrations. The first draft version was enhanced to also cover governance, culture and competencies and this version was used in this study as a basis.

In this customized data quality maturity measurement data quality maturity dimensions were evaluated in levels 1-4 with the following values included in the maturity assessment and heatmap:

Level 1: Red (as lowest score)
Level 2: Yellow
Level 3: Light green
Level 4: Green (highest score)

For each data quality maturity dimension, a corresponding score compared to the scoring criteria was given and updated in the heatmap. This way it is visible what are the DQ maturity dimensions which are already in good condition and which ones need to be improved. It is also possible to use this assessment tool to plan how to improve the data quality and what would be the most important DQ improvement actions and tasks to concentrate on for improving data quality maturity. This way, it would also be possible to prioritise and set targets for the DQ maturity dimensions.

This maturity tool was created for implementing regular measurements according to the need of the case company, i.e., every quarter, half of the year or yearly, to check the maturity and highlight improvements which had been achieved in the data quality maturity of a business-critical data asset in question. Further enhancements to the DQ maturity measurement tool had already been done according to the comments and feedback received from the pilot users of the case company. More

comments, feedback, and observations for improving the DQ maturity measurement tool will be expected once more measurements will be done so development of the tool for the case company will continue.

	Our data asset	Your chosen Dataset: 1) For example: data produced by <Org1>; consumed by <Org2>
TECHNICAL	**Data flow architecture** - where does data come from, where does it go to, which transformations are made to it	1
	Data familiarity, documentation - Do you generally know what the data you use is?	2
	Accuracy - data correctly represents reality	3
	Completeness - all expected (and required!) data values are present.	4
	Consistency - data is uniform across sources or over time	1
	Timeliness data is up to date and relevant.	2
	Relevance - data is applicable to the business question or problem being addressed.	3
	Validity - data is accurate and also represents what it claims to represent.	4
	Reliability - data represents what it claims to represent also over time – you can trust it to be valid also in the future, should you check	1
	Integrity - data is accurate, consistent, and complete	2
	Precision - The level of detail and specificity of data values.	3
	Format - The structure and syntax of data values is optimal for usage	4
	Duplication - Related to presence of duplicate records or data values.	1

Figure 18. Example of visualization of heatmap results when using all DQ maturity levels 1-4 with random evaluation results.

6.1 Technical DQ assessment

Technical DQ assessment part of the customized DQ maturity assessment is probably the most challenging part of the customized DQ assessment because it requires detailed knowledge about the data asset in question, data flow and technical DQ dimensions. Without detailed knowledge of the data asset and these DQ dimensions, the scoring of these DQ dimensions won't be trustworthy. During the study, it was noted the less familiar the pilot users were with the details of data asset, the more optimistically the score of the status was reported. In these cases, the reason for this could have also been the differences in the rating between the participants within a group with different opinions about the score leading to a compromised result as an average. So self-reporting is not to be taken as definitive measurement, but it is subjective. This highlights the fact that the scoring should be based on evidence, and if there is not enough evidence, these DQ dimensions should be assessed only once there is enough knowledge and evidence.

Data Quality Maturity Dimensions	Level 1 (Lowest)	Level 2	Level 3	Level 4
Data flow architecture - where does data come from, where does it go to, which transformations are made to it	Not even a rough idea where data comes from, by whom, via what, ending up where, responsibility lies on who?	Rudimentary knowledge of main data flows, but it is anecdotal/not documented - e.g. Architecture documents do not exist at all	Knowledge of main data flows exists and is documented - e.g. Architecture documents exist - all of these are how ever incomplete AND/OR not widely accessible / findable / competing version	Knowledge of main data flows exists and is documented - e.g. Architecture documents exist - and are complete and accessible for all
Our data asset	We don't know which datas are relevant or available to us or we are responsible for (...that other teams depend on)	We know the immediate inbound data that comes in front of us and we work with daily	We know our inbound data, we know who uses and depends on our data, we can easily point to documentation on either, we know who is responsible for what	New inbound data sources are communicated to our team, we also know of industry wide datasets we can enrich our own asset with; when a new party starts using our data, we know about it and can track it.
Data familiarity, documentation - Do you generally know what the data you use is?	No documentation available	Documentation is available, but not up to date or complete, or is disorganized, so it can not be fully used	Documentation is up to date and complete, but not widely used / people can not find it / master version of documentation is unclear or there are several competing masters	Documentation is up to date, complete for its purpose, widely used and easily accessible & found
Accuracy - data correctly represents reality	Data accuracy is unknown or not measured	Data accuracy is measured, but not consistently or with limited scope	Data accuracy is consistently measured for critical data sets, but not proactively improved	Data accuracy is consistently measured for all data sets and proactively improved (e.g. With automation)
Completeness - all expected (and required!) data values are present.	Data completeness is unknown or not measured	Data completeness is measured, but not consistently or with limited scope	Data completeness is consistently measured for critical data sets, but not proactively improved	Data completeness is consistently measured and monitored for all business relevant data sets and proactively improved
Consistency - data is uniform across sources or over time	Data consistency is unknown or not measured	Data consistency is measured, but not consistently or with limited scope	Data consistency is consistently measured for critical data sets, but not proactively improved, methodology not harmonized, no process.	Data consistency is consistently measured for all data sets and proactively improved with a clear and functioning process
Timeliness data is up-to-date and relevant.	Data timeliness is unknown or not measured	Data timeliness is measured, but not consistently or with limited scope	Data timeliness is consistently measured for critical data sets, but issues in it are not proactively improved	Data is continuously updated in real-time or near real-time as needed, with clearly defined processes for updating as needed
Relevance - data is applicable to the business question or problem being addressed.	Data relevance is unknown (see "Documentation")	There is silent knowledge/anecdotal trust in data being relevant to intended use	Partial trust in data relevance, which can be proven	Exact knowledge exists for what the data is suitable and not suitable for
Validity - data is accurate and also represents what it claims to represent.	Data validity is unknown or not measured **(see "Data familiarity, documentation" - if you don't have any documentation, how would know whether or not.)**	There is silent knowledge/anecdotal trust in data being valid to intended use	Partial trust in data validity, which can be proven	Exact knowledge exists for what the data is suitable and not suitable for
Reliability - data represents what it claims to represent also over time – you can trust it to be valid also in the future, should you check	Data reliability is unknown or not measured	Data reliability is measured, but not consistently or with limited scope	Data reliability is consistently measured for all data sets and proactively improved	Advanced technologies and techniques are used to continuously monitor and improve data reliability
Integrity - data is accurate, consistent, and complete	Data integrity is unknown or not measured	Data integrity is measured, but not consistently or with limited scope	Data integrity is consistently measured for critical data sets, but not proactively improved, does not have a process	Advanced technologies and techniques are used to automatically monitor and improve data integrity
Precision - The level of detail and specificity of data values.	Data precision is unknown or not measured	Data precision is measured, but not consistently or with limited scope	Data precision is consistently measured for critical data sets, but not proactively improved	Advanced technologies and techniques are used to continuously monitor and improve data precision
Format - The structure and syntax of data values is optimal for usage	Data formats just are, no thought is given to whether the format is useful, unuseful or actively dysfunctional	Data format is known, but local ungeneralizable quirks and habits are all around the data	Data format is suitable for use, but not of a type future-proof or viable standard	Data formats are useful and futureproof (e.g. established standards)
Duplication - Related to presence of duplicate records or data values.	No knowledge of duplicates, no capability to recognize them.	Knowledge of duplicates, capability to recognize them, but not much done about it	Duplicates found and dealt with, but e.g. Manual post - hoc corrections without a process	Automated recognition of duplicates and fixing it.

Figure 19. Technical data quality dimensions including the definitions of their scores in the first version of the customized data quality maturity tool.

6.2 Governance DQ assessment

Governance part of the DQ assessment include these DQ dimensions which are fundamental for improving data quality. These require actions which have to be taken universally across the organization, i.e. to be included in the business processes. It was noted during the study that the score

for the data quality metrics DQ dimension was identical to all groups in a way that even though some metrics are in place and data quality is monitored, there is room for improvement and creating more DQ metrics.

Data Quality Maturity Dimensions	Level 1 (Lowest)	Level 2	Level 3	Level 4
Data quality governance	DQ just is what it is. Not much is done about it	Processes are planned / some exist but they are not uniform, not used. No improvements are planned.	A data quality improvement lifecycle process in which DQ issues are recognized, collected, prioritized, analyzed, improved, controlled has **been planned** but is not totally in use.	Working data quality improvement lifecycle process in which DQ issues are recognized, collected to a single accessible place, issues are prioritized according to business value, analyzed, improved, and (automatically) controlled **is in use**
Data quality standards	No standards, not within team, not across organization	Standard within team, not across organization	Standards matching across teams for suitable parts has been planned	Standards matching across teams for suitable parts has been taken into use. **Data is validated accordingly in the source system(s) - downstream doesn't need to worry.**
Data quality metrics	DQ is not logged, followed, measured or reported.	Some data metrics are used for some parts of a datasets dataflow (ETL - phases)	Most critical datasets/parts are monitored from ALL phases, where changes are made (e.g. extract, transform, load).	Critical datasets/parts are monitored automatically, a process exists and is enforced where new datasets are included as parts of metrics you can trust to depict data features in a trustworthy manner.
DQ issue reporting and mitigation	No one to report a DQ issue to, you find it, you own fixing it (or forget it). No processes, no help, you are on your own	There is at least a group you can consult to get help with	If you find an issue, you know where to report it to and it will be dealt with, Root causes of the DQ issues are analyzed and corrected whenever possible.	DQ issues are found and logged automatically and fixed automatically. State of the art: learning neural networks.
DQ processes	No process for anything. Someone maybe does something. Or maybe not.	Are processes in place for performing root-cause-analysis for new data quality issues?	Are at least some DQ problems dealt with proactively (before they happen)? Are there master data and reference master data policies and automated checks in place so no duplicate data is made?	Critical datasets/parts are monitored automatically and a violation launches a DQ mitigation process automatically. The success and progress of these processes themselves are measured
General data management	Data is not managed, it just comes and is used in a haphazard manner, where information is left as silent knowledge in the project participants minds and dissipates as team members leave or start doing something else. Data quality and its features are unknown.	There is a process of including existing and incoming new datasets as part of a data asset listing, with documentation and its features profiled, at least within business line.	Aforementioned process is enforced and data with its features is included into a data catalogue on a regular basis.	Aforementioned process is enforced and data with its features is included into a data catalogue on a regular basis at least partly with automation, so when new data comes in (e.g. via a company merger) you can trust it to have documentation enabling its use available quite soon.

Figure 20. Governance related data quality dimensions including the definitions of their scores in the first version of the customized data quality maturity tool.

6.3 Culture and Competencies DQ assessment

Culture and Competencies DQ dimensions were added to the first version of the customized data quality maturity assessment to bring out the importance of the collaboration and competence development which are required to accelerate the data quality improvement actions. These kind of DQ dimensions should be understood to be vital for the DQ maturity improvement, in addition to Governance related DQ dimensions.

CULTURE & COMPETENCIES	**Collaboration: Silo / Group work**	No cross-team collaboration nor knowledge-sharing exists for handling data quality issues and developing data quality practices and competencies. Data producer gets no information about data quality issues affecting data consumers work. Working in silos.	Some teams/team members (Data Producers and Data Consumers) collaborate and share knowledge with each other for informing & handling data quality issues and developing data quality practices and competencies but it is ad-hoc. No processes or ways of working exist.	Some teams/team members (Data Producers and Data Consumers) regularly collaborate and share knowledge with each other for informing & handling data quality issues and developing data quality practices and competencies. Processes or ways of working exist.	Cross-team collaboration and knowledge sharing is in practice across organization in order to report & handle data quality issues and developing data quality practices and competencies. Clear RACI (Responsibility, accountability, consulted, informed) matrix exists for BOTH data producer AND consumer, it is used and adhered to and updated as needed.
	Self-service capabilities	No possibilities to interact with the data nor create data quality monitoring reports. Data is a black box (for example in the case data producer is external)	Only IT and developers are able to interact with the data and create & share data quality monitoring reports IF they have time and resources - no SLA, no possibility to create self-service reports.	Capabilities exist for independently interacting with the data but those are not used for creating self-service data quality monitoring reports. (E.g. DQ tool is offered, connecting to dataset possible, but no-one uses them.)	Capabilities exist for independently interacting with the data and create self-service data quality monitoring reports and the ways of doing so are familiar to more than a single person in organization.

Figure 21. Culture and Competencies related data quality dimensions including the definitions of their scores in the first version of the customized data quality maturity tool.

7 PREREQUISITES FOR THE CUSTOM DQ MATURITY ASSESSMENT

The first task for the pilot users of the customized DQ Maturity Assessment was to choose a data asset for which the DQ maturity assessment would be done. Instructions for this were to choose a business-critical data asset. For the chosen data asset, it was also requested to write down whether it is something they use as data consumer, or which they produce for others to use i.e. act as data producer (Figure 22). That was to highlight the roles of data producer and data consumer, and whether for example data producers have knowledge of their data consumers, and vice versa.

Data asset - EXERCISE SCOPE: choose the business-critical data asset you handle in this exercise and write down whether it is something you use or something you produce for others to use.	**Your chosen Dataset:** produced by WHO?; consumed by WHO?

Figure 22. Instructions for choosing the business-critical data asset for custom DQ maturity assessment.

While doing this study, it was observed there are certain prerequisites for doing this custom DQ Maturity Assessment. These were not included in the DQ Maturity assessment tool as pre-requisites but during this study, it became clear that to benefit the business in the best possible way, these prerequisites should be included in the future.

7.1 Business critical data asset

Selecting a business-critical data asset for the assessment. It should be of high value to the business to get as much benefit as possible of the assessment. Business critical data assets are the data assets which are the most important one related to an organization's business operations, decision making, customer's needs and demands for fulfilling regulatory and compliance requirements. However, each organization must themselves decide which data they consider to be critical for them and making these decisions should be led by business. Business critical data asset can even be for example one column, like ID, if that is critical for the business and must be consistent in every system and system integrations. Additionally, it should be taken into account that selecting the business-critical data asset is not a department specific decision to be done in silos, but it should be based on the core business purposes of each organization to cover the business operations and decision making end-to-end.

Examples of the business-critical data assets are master data like:

- Customer data
- Product/Material data
- Employee data
- Vendor data

Transactional data like:

- Operational data
- Financial data required for auditing purposes
- Corporate financial data

However, these are only high-level examples. In the details of these data assets, there are data assets which add more value to the business than others. So in the end, the data quality activities should be targeted even to the detail level of these high level, business-critical data assets. For example, in the worst-case scenario, it can be possible to be fined for even one PII in the wrong field so it should be very carefully to be monitored and followed-up this kind of case does not arise (GDPR, 2020).

Regarding the decision-making process which is based on the data, reports and dashboards delivered by data team, there is a different kind of approach to find out the business-critical data.
For example, in his article, Dengsøe gives practical steps for identifying business-critical data from the data team perspective, like data models and dashboards (Mikkel Dengsøe, 2023). He states it is important to identify these business-critical downstream dependencies and use cases. These can be tagged critical so there would be visibility about these critical data assets which effect to this critical dashboard, like in the Figure 23. In the data team case, the business-critical data asset for the DQ assessment could be for example a very important dashboard. For this business-critical dashboard, all these data models upstream are on the critical path (Figure 23). Data flow of this dashboard should be the first task to find out and document, preferably in the data catalogue. In addition to the reports and dashboards, the data which has to be reported to the regulatory authorities is critical. The reason for this is C-suite can be held personally liable for reporting incorrect data to regulators (Dengsøe, Murphy 2023). Additionally, now that the machine learning and artificial intelligence implementations are increasing, it is important to get transparency of the data and

data flow effecting to these. For this purpose, a business-critical implementation could be selected and the data effecting to this should be tracked before doing the DQ maturity assessment. Data and data lineage of artificial intelligence and machine learning implementations will also be regulated soon so this will increase the business-criticality of these (European Commission, Joint Research Centre, Balahur, Jenet, Hupont Torres et al, 2022). Although it is a risk-based model, and it will not cover all the development, it will lead the way for inclusive, non-biased AI/ML development. Consequently, it increases the importance of the DQ maturity assessments of the data which is used as basis of these AI/ML implementations.

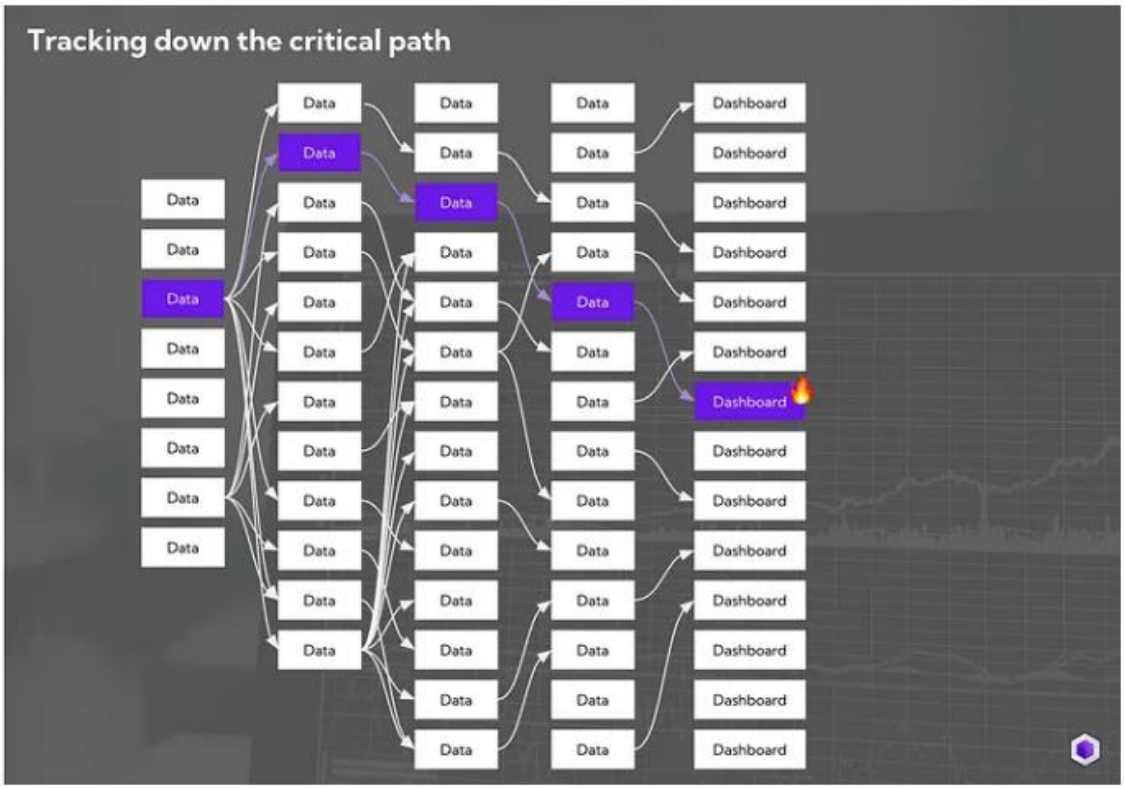


Figure 23. Tracking down the critical path for identifying business-critical data from the data team perspective (*Dengsøe, Murphy 2023).*

7.2 Data literacy of the business-critical data asset to be assessed

According to Sebastian-Coleman, "*Data literacy is the ability to read, understand, interpret, and learn from data in different contexts and to communicate about data with other people"*. It requires skills, knowledge, and experience to become data literate. Data literacy of the business-critical data asset is a must for doing the data quality maturity assessment for it and get the most benefit out of the assessment. Data literacy in this context means the selected data asset should be familiar

enough to the one(s) who do the assessment. However, it is one kind of result if it turns out the selected data asset is not familiar enough and it would mean the data asset in question should first be familiarized with and then do the assessment again. Note it is possible to use this DQ maturity assessment both individually and as a groupwork. When doing it as a a groupwork, everyone included in the group will not need to be on the same level of data literacy, but everyone has a chance to learn from each other. It is also important to note, according to Sebastin-Coleman,

> "*No single individual can know everything about an organization's data. But, together, people can solve more problems in better ways if they understand data as a construct, recognize the risks associated with data production and use, cultivate a level of skepticism about data, and develop skill in visualizing and interpreting data. They will solve even more problems if the organization supports these efforts through disciplined metadata management and data quality management."*
> *(Sebastian-Coleman, 2022)*

7.3 Data flow, data catalogue and data lineage

For getting the most benefits of the data quality maturity assessment, the data flow architecture should be known and documented. This kind of documentation includes the information about where the data comes from, where it goes to, and which transformations are made to it in between. Data flow should be documented transparently, and it should be easily found and accessible by the data consumers. Easily accessible data documentation accelerates the usage of the data in the implementations in reporting, automation, artificial intelligence, and machine learning, and ensures the knowledge is not lost due to changes in personnel. Preferably this kind of documentation is found and visible in the data catalogue. Data catalogue includes metadata type of information about for example source systems of organization, datasets, data owners, data/business stewards, business glossary, data lineage, reports, and data quality.

For example, in Collibra Data Catalogue tool, there are two types of data lineages:

1) **Technical lineage** (Figure 24): more detailed data flows including data objects like tables and columns, data transformations and source code for information to e.g., data engineers, data architects and technical stewards.

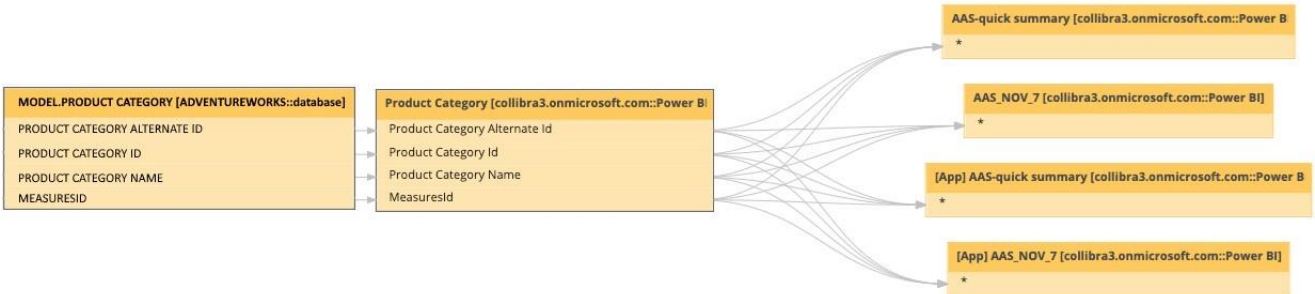


Figure 24. Example of technical lineage by Collibra Data Catalogue tool (Collibra, 2023).

2) **Business lineage** (Figure 25): a summary of technical lineage for business stewards and analysts

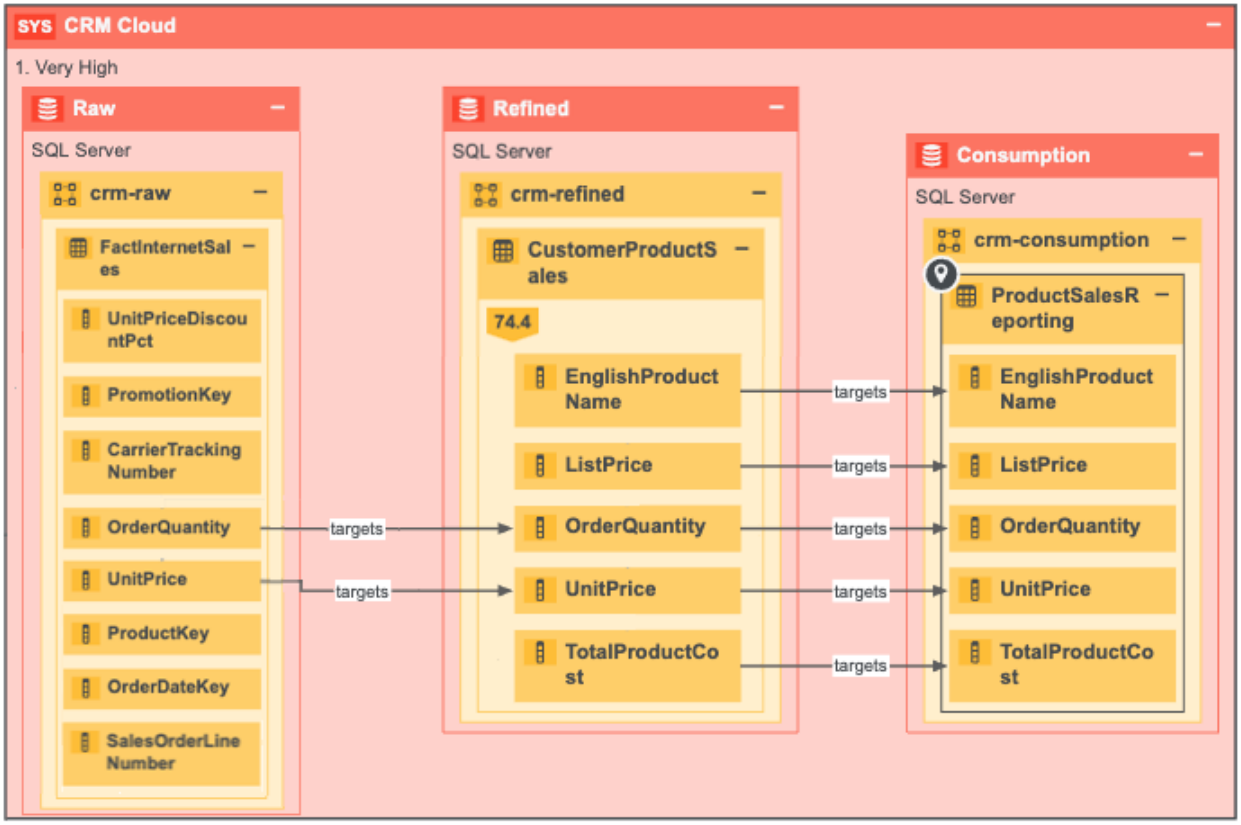


Figure 25. Example of business lineage by Collibra Data Catalogue tool (Collibra, 2023).

Via data lineage it is possible to have auditable trail of data transformations across the whole data flow. Data quality should be considered and monitored in all parts of the data flow (Figure 26). By using data catalogue tool, it is also possible to create visibility to the data quality rules and metrics of the data assets and automate the data issue creations, so these are linked to the data asset in question. From the end-to-end data lineage, it is possible to see where this data issue has effect to.

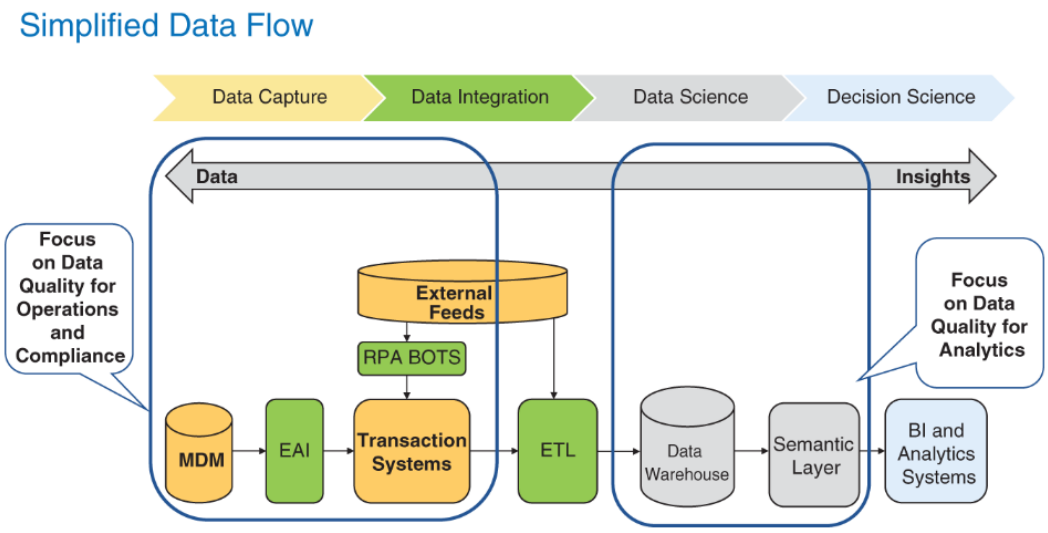


Figure 26. Simplified Data Flow (Southekal, 2023).

8 IMPROVING DATA QUALITY MATURITY

First step for improving data quality maturity is to prioritise data quality improvement and analyse the current state of DQ maturity by using a DQ maturity model. Based on the current state analysis, it is possible to plan the concrete DQ improvement actions in more detail according to the business needs of your organization. After doing the current state analysis, setting the targets for improvements, and prioritising these actions which can contribute to achieving the next levels in DQ maturity, and with the business benefits always in the focus. For the concrete improvement activities, even quite detailed level instructions, possible templates, guidance, and leadership are needed. The most effective improvement actions are the ones which effect to the root causes of the data quality issues.

To set realistic goals for the DQ improvement actions, the required resources, their competencies, needed roles and responsibilities for achieving these goals must be taken into account. Without competent enough resources with properly defined roles and responsibilities, only ad-hoc improvements in DQ maturity will be achieved. DQ improvements are not something one can do without dedicated time for it. Unfortunately, this is often neglected because the financial benefit for this work is not directly visible but considered as a cost for an organization. According to Jones (2017), the iceberg of ignorance is evident in the data quality context as well, meaning that only 4% of the data quality problems are known to the top management and only 9% of these are known to middle management. Most of the data quality problems are below the visibility of the management and only known by the supervisors and employees (Jones, 2017). Therefore, it is vitally important the management is aware of these invisible costs. Better yet would be to start to collect and follow up the costs of doing the corrections of poor data, which in the worst-case scenario, are done manually and even by subcontractors. To get visibility to these costs would be one way to get buy-in and support from the management for the data quality activities. Improving data quality maturity requires concrete actions for creating data quality metrics and targets for these. These metrics must be regularly monitored, and the DQ maturity regularly assessed to find out and share the progress of DQ improvement actions.

9 ANALYSIS & RECOMMENDATIONS

9.1 Results

Research questions of this study were:

1) Is this kind of custom data quality maturity assessment tool useful in improving data quality of an organization in a systematic way?
2) How does this custom data quality maturity assessment tool compare to the existing data quality maturity models?

Hypothesis of this study was using custom data quality maturity assessment tool helps the organization to get more awareness, understanding and help them to plan actions to improve data quality and data quality maturity in general. Additionally, it can also provide transparency to the current state of the data quality maturity which is very important especially for the business-critical data assets. In the case company this kind of more detailed data quality maturity model proved its purpose because, in addition to the data quality trainings, it has been decided to continue the usage and further development of this kind of customized data quality maturity assessment in the case company. Since it typically takes time to improve data quality so it would be visible even by using in this customized, more detailed data quality maturity assessment, there was not enough time for gathering evidence about improved data quality during this research. In addition, the improvement of the awareness and understanding of data quality is quite challenging to measure. It was not measured during this research so these results are not available as a result of this study. One kind of measurement for the improvement of understanding of data quality of business-critical data assets could have been for example the addition of data profiling practises to the processes but there was not enough time and resources to introduce these kinds of new practices during this study. However, according to the observations, the data quality awareness, and understanding was improved in the case company during this research, and it has led to the planning of the concrete data quality improvement actions of the business-critical data assets. There is also a plan to start and experiment using this custom data quality maturity tool for KPI status, target settings and continuous measurements. KPIs are to be calculated based on the results of the DQ dimension scores.

During this study, no other comparable data quality maturity measurement assessment was found; instead, all the DQ maturity measurements assessments which could be studied, are in too high a level to be operationally useful. With reference to this, it was found the data quality maturity measurements are not comparable even though it was not possible to compare all the available methods because most of these must be purchased. Additionally, because data quality mostly depends on data management and data governance, data quality is often just one part of these data management and data governance maturity measurements. In addition, it was noted this first version of the custom data quality maturity tool also requires updates to enhance and accelerate the usage of it in the case company. Consequently, feedback and observations for improving the customized data quality maturity template have been considered and implemented simultaneously while doing this study. One important improvement was to include "how-to" instructions for the DQ dimensions to guide the users to get to the next level of DQ maturity. It was also noted that this custom data quality tool might be in too detailed level for the higher management, so it is under planning whether there is a need to create another version for the management, or use another, higher level DQ maturity model for the management. Another idea for improving the custom DQ maturity model is to include the responsible roles/owners for each DQ dimension but this would also require the higher-level ownership to get the resourcing and budget for these data quality improvement actions.

9.2 Discussion

It has been noted during this study that this customized data quality maturity template is created and improved in such a way it could be used also in other organizations either for data quality training and/or improving data quality awareness and visibility. It can also be used for planning data quality improvement actions and for measuring the effect of these improvement actions in quite a detailed level. Even KPIs and targets regarding these could be created based on the current state of the data quality maturity.

Limitations for the usage of this kind of customized data quality maturity assessment is that the data quality related details which are included in the assessment, might not be self-evident for these who are not so familiar with the details of the business-critical data asset in question, and the data quality terminology. So even though this customized DQ maturity assessment tool has already been enhanced in the case company, it still requires further improvement so it can be understood by such a user who does not have enough knowledge to do self-evaluation of all these questions

which are included in the customized DQ maturity assessment tool. In such cases, it cannot be used effectively without guidance and discussion about the meanings of the DQ dimensions and scores. This would need to be studied more and further enhancements to the maturity tool accordingly.

In addition, it was noted this kind of self-evaluation without evidence probably gives more optimistic results than what the actual status is. To get reliable results of the assessment, there would need to be evidence as reference included in the heatmap, for example, in the comments of the heatmap field. This would be useful for information sharing purposes as well. Also, to get some improvements to the data quality maturity, it would be more useful to give a lower score than a more optimistic one. In case there is not enough knowledge of the actual status, it would be skipped, and score(s) would be given only once there is enough knowledge and/or evidence, of the actual status.

Management support is vital for improving data quality since the DQ improvement actions are not taken by accident. However, this kind of more detailed data quality maturity assessment might be too detailed for the management to understand the data quality and the requirements for improving the data quality. It might be there is a need to have a separate, higher level data quality maturity assessment for the management? It would have to be linked to this more detailed one. It could be for example that only some parts of the DQ maturity assessment are included in the management part of the DQ assessment, and the target for the more detailed DQ assessment would be set according to these. Also, if the management is familiar with some other kind of higher level DQ maturity assessment, they could be linked to get visible improvements also in the higher level DQ maturity assessment.

9.3 Reliability and validity of research

The effect of the usage of this customized data quality maturity template was experimented in only one case company so it should be experimented in some more organizations. This would give more information and feedback about the usage and contribution of using this kind of maturity tool for improving data quality and related practices within different kinds of organizations. In the literature review, no other comparable data quality maturity models were found but not all these maturity models were studied in this thesis. Most of data quality and data governance maturity models would

have had to be purchased, and, since the maturity models are not standardized, they are not comparable either.

Additionally, the actions needed for improving data quality typically take time and resources to implement. This study should have taken more time for the data quality improvement to be noted in the assessment. However, while doing this study it has proved its purpose for the case company to use customized and more detailed data quality maturity model for assessing organization's data quality maturity and use this information for planning what kind of actions are needed to improve data quality. It was also found it is possible to set the KPI targets after analysing the current state of DQ maturity and follow-up the progress even with just one KPI target number, aggregated from the DQ maturity tool. This way, including the regular DQ maturity assessments, it can be used as a continuous practice which also enables to develop the maturity model and improve data quality practices according to the need of the organization in question. For the management, it can offer more detailed transparency about the data quality maturity of the company in the business-critical data asset level.

9.4 Further research

The first idea for further research would be to use this kind of custom data quality maturity tool in another organization, or even several different organizations to get their results of using it to compare the results of this thesis.

Further research would also be needed to create more automation into the data quality maturity assessment and achieve automated visibility about the data quality and data quality maturity of the business-critical data assets in an organization. Continuous data quality monitoring and automated corrections of business-critical data assets, wherever possible, must become business as usual in near future. Column-level validations of business-critical source data, monitoring the integrations between systems and ensuring the consistency of the data between systems could solve most challenges in data quality.

In addition, it would be feasible to study the AI systems' data quality related aspects in more detail and develop a maturity model for assessment. It should be done to ensure publishing transparent and trustworthy AI systems where their development includes all these mandatory aspects for

inclusive and non-biased AI which will be required by the EU AI Act soon (European Commission, Joint Research Centre, Balahur, Jenet, Hupont Torres et al, 2022). There is now an urgent need for the organizations to prioritize their efforts to address data quality challenges, ensuring their AI systems deliver accurate, reliable, and unbiased results. For this purpose, MS Excel and self-evaluations of data quality maturity are not enough to prove it without automated monitoring and evidence.

10 CONCLUSIONS

Improving data quality is not rocket science. It requires concrete actions which are to be done end-to-end in good collaboration with the specialists in IT, data management, and especially with the business specialist who are experts in the business-critical data assets and these details which have an effect to business, and consequently, the bottom line. It should be a default, and embedded into the business processes, to measure the data quality maturity as a starting point, and while doing this, plan, set priorities and targets for the improvement actions. This way targets should be followed-up across the organization so the improvements of the data quality maturity would be visible. This would enable to also share learnings, best practices and improvement actions which contributed to the improved DQ maturity across the entire organization.

Data quality related knowledge and understanding seems to be challenging and time consuming to achieve. It requires attention to detail but also business understanding about the data in question and how it relates to the big picture from the business point of view. However, when given the concrete visibility to the data, and the related information and/or documentation about it, everyone can learn to “read data”. Nowadays it is a must to learn, and continuous learning is everyone’s responsibility, just like the data quality improvement is. People, their knowledge and experience about business and data, data literacy competencies, and last but not least, collaboration and collective intelligence are the keys to the data quality improvements. This kind of culture requires individuals to learn and collaborate across the whole organization, and the organizations to prioritise and allocate resources to data related learning and concrete actions to improve data quality. Devil is in the details, and nobody can by themselves know every detail which effect to data quality and can have significant effect to the business. Business-criticality of the data should be visible and known to all employees end-to-end of the business processes. This way the actions and resources can be prioritised to these data assets and their details which are the most important from the business point of view.

ACKNOWLEDGEMENTS

I'd like to thank various people who have contributed to this study and thesis. First, I want to thank my supervisors for the opportunity to do this study about experimenting and piloting customized data quality maturity tool, Matti Laamanen and my colleagues for support and valuable comments. Then, extra special thanks to consultant, PhD Manne Laukkanen who did the first draft version of the customized data quality maturity tool as a basis for the DQ training and continued to develop it further in collaboration and based on feedback. Finally, many thanks to PhD Virtanen, my master thesis supervisor, for his professional guidance, very useful feedback and recommendations.

REFERENCES

alexbarros, Github, ydataai / ydata-profiling, 2023. Search date 3.10.2023 GitHub - ydataai/ydata-profiling: 1 Line of code data quality profiling & exploratory data analysis for Pandas and Spark DataFrames.

Baskarada, Sasa, 2009. Chapter 1.3 Justifications for the research, Figure 1.1 The Impact of Poor Information Quality, Source: developed from Redman, 1998. Search date 26.10.2023 (PDF) Information Quality Management Capability Maturity Model (researchgate.net)

Baskarada, Sasa, 2009; Table 2.23: Bitterer A, Gartner's Data Quality Maturity Model, Gartner Research, 2007. Search date 20.10.2023 (PDF) Information Quality Management Capability Maturity Model (researchgate.net)

Belghith, Oumaima, Sirine Zitoun, Sabri Skhiri dit Gabouje, Syrine Ferjaoui, 2021. A Survey of Maturity Models in Data Management. Search date 10.10.2023 (PDF) A Survey of Maturity Models in Data Management (researchgate.net)

Chang, Robert, Airbnb Tech Blog, 2021. Search date 30.11.2023 How Airbnb Achieved Metric Consistency at Scale How Airbnb Achieved Metric Consistency at Scale | by Robert Chang | The Airbnb Tech Blog | Medium

Collibra, 2023. Collibra Data Lineage. Search date 15.11.2023 Collibra Data Lineage

Collibra, 2023. High quality data is the foundation of ESG. Search date 22.10.2023
High quality data is the foundation of ESG | Collibra

DAMA UK Working Group on Data Quality Dimensions, 2013. The Six Primary Dimensions for Data Quality Assessment. Search date 30.9.2023
https://www.sbctc.edu/resources/documents/colleges-staff/commissions-councils/dgc/data-quality-deminsions.pdf

DataCamp, 2023. Datacamp Data Quality Dimensions, Introduction to Data Quality. Search date 21.10.2023 *Data_Quality_Dimensions_yebbuu.pdf (datacamp.com)*

Deloitte, 2022. Augmented Data Management: Beyond the Hype. Search date 30.10.2023 Augmented Data Management: Beyond the Hype | Deloitte Netherlands

Dengsøe, Mikkel, Murphy, Lindsay, 2023. How to Identify Your Business-Critical Data. Search date 21.10.2023 How to Identify Your Business-Critical Data | by Mikkel Dengsøe | Towards Data Science

Ehrlinger Lisa and Wöß Wolfram, 2022, A Survey of Data Quality Measurement and Monitoring Tools. Frontiers Big Data. Search date 30.9.2023
Frontiers | A Survey of Data Quality Measurement and Monitoring Tools (frontiersin.org)

European Commission, Joint Research Centre, Balahur, Jenet, Hupont Torres et al, 2022. Data quality requirements for inclusive, non-biased and trustworthy AI. Search date 1.10.2023
Data quality requirements for inclusive, non-biased and trustworthy AI - Publications Office of the EU (europa.eu)

Experian, 2023. Discover and profile data. Search date 13.11.2023
Data Quality user documentation | Discover and profile data (experianaperture.io)

Experian, 2023. Experian's Data Maturity Assessment. Search date 1.10.2023 Data Quality Maturity Assessment (experian.co.uk)

Experian, 2023. What is a Data Quality Standard? Search date 25.10.2023.
What is a Data Quality Standard? | Experian Business

Forbes, 2022. The Journey To Augmented Data Quality. Search date 30.9.2023 The Journey To Augmented Data Quality (forbes.com)

Forbes, Kelly Anne Smith 2022. Greenwashing And ESG: What You Need To Know. Search date
Greenwashing And ESG: What You Need To Know – Forbes Advisor

Gartner, 2007. Gartner's Data Quality Maturity Model. Seach date 1.10.2023 Gartner's Data Quality Maturity Model

Gartner, 2018. How to Stop Data Quality Undermining Your Business. Search date 30.9.2023 4 Steps to Overcome Data Quality Challenges (gartner.com)

Gartner, 2021. Search date 4.10.2023 12 Actions to Improve Your Data Quality (gartner.com)

Gartner, 2022. Magic Quadrant for Data Quality Solutions. Search date 8.10.2023 Gartner Reprint

Gartner, Peer Insights, 2023. Data Quality Solutions (Transitioning to Augmented Data Quality Solutions) Reviews and Ratings. Search date 22.10.2023 Best Data Quality Solutions Data Quality Solutions (Transitioning to Augmented Data Quality Solutions) Reviews 2023 | Gartner Peer Insights

Gawande, Sandesh 2022. 6 Dimensions of Data Quality, Examples, and Measurement. Seach date 30.9.2023 A Guide for Data Quality (DQ) and 6 Data Quality Dimensions (icedq.com)

GDPR, 2012. Personal data protection: processing and free movement of data (General Data Protection Regulation). Search date 30.10.2023 2012/0011(COD) - 27/04/2016 - Personal data protection: processing and free movement of data (General Data Protection Regulation) (europa.eu)

GDPR, 2020. Complete guide to GDPR compliance. Search date 23.10.2023 General Data Protection Regulation (GDPR) Compliance Guidelines

GDPR, 2020. How the GDPR could change in 2020. Search date 22.10.2023 2020 developments for data protection and the GDPR - GDPR.eu

Great Expectations, 2023. Great Expectations. Search date 3.10.2023 Welcome | Great Expectations

Grepsr, 2021. Perfecting the 1:10:100 Rule in Data Quality. Search date 20.10.2023 Data Quality Ensured: Perfecting the 1:10:100 Rule | Grepsr

Hao, Karen, 2019. Training a single AI model can emit as much carbon as five cars in their lifetimes. Search date 22.10.2023 Training a single AI model can emit as much carbon as five cars in their lifetimes | MIT Technology Review

IBM, 2023. IBM SPSS Modeler. Seach date 29.11.2023 IBM SPSS Modeler

Intozetta, 2023. Data Quality - IntoZetta

Jones, Clinton, 2017. The iceberg of data quality ignorance. Search date 30.9.2023. The iceberg of data quality ignorance (linkedin.com)

ISO 25012, 2022. ISO 25000 software and data quality, ISO/IEC 25012. Search date 23.10.2023 ISO 25012 (iso25000.com)

ISO 3166, 2020. ISO 3166 Country Codes. Search date 24.10.2023 ISO - ISO 3166 — Country Codes

ISO 4217, 2015. Search date 2023 ISO - ISO 4217 — Currency codes

Marr, Bernard, Forbes, 2023. Green Intelligence: Why Data And AI Must Become More Sustainable. Search date 22.10.2023 Green Intelligence: Why Data And AI Must Become More Sustainable (forbes.com)

McCallum, Q., Ethan, 2012. Bad Data Handbook.

Maydanchik, Arkady, 2012. Data Quality Assessment.

Oval Edge, 2023. Data Governance Maturity Models and How to Measure It? Search date 3.10.2023 Data Governance Maturity Models and How to Measure It? (ovaledge.com)

Politico, 2022. Dutch scandal serves as a warning for Europe over risks of using algorithms. Search date 30.10.2023 Dutch scandal serves as a warning for Europe over risks of using algorithms – POLITICO

Quoss, Vaughn, Airbnb Tech Blog, Medium, 2020. Data Quality at Airbnb. Search date 29.11.2023 Data Quality at Airbnb. Part 2 — A New Gold Standard | by Vaughn Quoss | The Airbnb Tech Blog | Medium

Redman, Thomas C., 1998. The impact of poor data quality on the typical enterprise. Search date 30.9.2023 The impact of poor data quality on the typical enterprise | Communications of the ACM

Redman, Thomas C., Harvard Business Review, 2016. Data quality should be everyone's job. Search date 30.9.2023 https://hbr.org/2016/05/data-quality-should-be-everyones-job

Redman, Thomas C., Harvard Business Review, 2018. If Your Data Is Bad, Your Machine Learning Tools Are Useless. Search date 30.09.2023 If Your Data Is Bad, Your Machine Learning Tools Are Useless (hbr.org)

Redman, Thomas C., Harvard Business Review, 2020. To Improve Data Quality, Start at the Source. Search date 30.9.2023 To Improve Data Quality, Start at the Source (hbr.org)

Redman, Thomas C., Harvard Business Review, 2022. Bad Data Is Sapping Your Team's Productivity. Search date 30.9.2023 Bad Data Is Sapping Your Team's Productivity (hbr.org)

Redman, Thomas C., Basecap Analytics, 2022. Is Poor Data Quality Holding You Back? Search date 30.9.2023 Is Poor Data Quality Holding You Back? - Basecap Analytics

RenChu Wang and dorisjlee, Github, lux-org / lux, 2023. Search date 3.10.2023 GitHub - lux-org/lux: Automatically visualize your pandas dataframe via a single print!

ScienceDirect, 2021. ISO 4217 Currency codes. Search date 21.10.2023
Data Quality Dimension - an overview | ScienceDirect Topics

Sebastian-Coleman, L., 2013. Measuring Data Quality for Ongoing Improvement:
A Data Quality Assessment Framework. Waltham, MA: Elsevier. Search date 1.10.2023
Measuring Data Quality for Ongoing Improvement - 1st Edition (elsevier.com)

Laura Sebastian-Coleman, 2022. The People Challenge: Building Data Literacy, Chapter 7. Search date 30.9.2023 The People Challenge: Building Data Literacy - ScienceDirect

Southekal, Prashanth, 2022. Root Cause Analysis (RCA) for Effective Data Quality Management. Search date 30.9.2023. Root Cause Analysis (RCA) for Effective Data Quality Management (uarrow.cloud)

Southekal, Prashanth, 2023. Data Quality, Chapter 9: Best Practices to Realize Data Quality. CHAPTER 9: Best Practices to Realize Data Quality | Data Quality (oreilly.com)

Statista, 2023. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025. Search date: 30.11.2023 Data growth worldwide 2010-2025 | Statista

Steenbeek, Irina, 2021. DAMA-DMBOK2 vs DCAM® 2.2: Maturity Models and Assessment. Search date 30.9.2023 DAMA-DMBOK2 vs DCAM® 2.2: Maturity Models and Assessment - Data Crossroads

Strubell, Emma, Ananya Ganesh, Andrew McCallum, 2019. Energy and Policy Considerations for Deep Learning in NLP. Search date 22.10.2023 [1906.02243] Energy and Policy Considerations for Deep Learning in NLP (arxiv.org)

Statistics Finland, 2023. Price index of electricity corrected in the Harmonised Index of Consumer Prices; also has an effect on total inflation. Search date 3.10.2023
Price index of electricity corrected in the Harmonised Index of Consumer Prices; also has an effect on total inflation | Statistics Finland

taylorfturner, Github, capitalone / DataProfiler. Search date 3.10.2023 GitHub - capitalone/DataProfiler: What's in your data? Extract schema, statistics and entities from datasets

TDWI, 2023. TDWI Data Management Maturity Model Assessment. Search date 10.10.2023 TDWI Data Management Maturity Model Assessment | Transforming Data with Intelligence

Uber, 2023. D3: An Automated System to Detect Data Drifts. Search date 30.9.2023 D3: An Automated System to Detect Data Drifts | Uber Blog

UK Government Data Quality Framework, 2020. The Government Data Quality Framework Seach date 23.10.2023. The Government Data Quality Framework - GOV.UK (www.gov.uk)

Van Nederpelt & Black, 2020. Dimensions of Data Quality (DDQ), Chapter 3.5 Result step 5: Summarize the results. Search date 30.9.2023 Code for Information Quality 2019 (dama-nl.org)

Victor Schmidt, Alexandra (Sasha) Luccioni, Alexandre Lacoste, Thomas Dandres, 2023. Machine Learning Emissions Calculator. Search date 22.10.2023 Machine Learning CO2 Impact Calculator (mlco2.github.io)

Wright, Clark, Airbnb Tech Blog, Medium, 2023. Data Quality Score: The next chapter of data quality at Airbnb. Search date 30.11.2023 Data Quality Score: The next chapter of data quality at Airbnb | by Clark Wright | The Airbnb Tech Blog | Nov, 2023 | Medium

Zhu Hongwei, Madnick Stuart E., Lee Yang W., Wang Richard Y., 2012. Data and Information Quality Research: Its Evolution and Future. Search date 29.11.2023
Madnick_2012_Data and Information Quality.pdf (mit.edu)