UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Karen Roht

Generative AI in Data Quality Management

Bachelor's Thesis (9 ECTS)

Supervisor: Anastasija Nikiforova, PhD

Tartu 2025

Generative AI in Data Quality Management

Abstract:
High-quality data is essential for making informed and accurate decisions, yet achieving and maintaining the quality of data can be a complex and resource-intensive challenge. The recent emergence of generative AI has the potential to facilitate data quality management and minimise manual work. This thesis explores how generative AI could be used in data quality management to address these challenges. The thesis proposes a typology, developed based on the systematic analysis of data quality tools available on the market, which was then validated and refined with experts. Only 13 of the 88 reviewed tools currently integrate generative AI. While experts saw potential in automating routine tasks and improving scalability, concerns were raised around trust, explainability, and data security. Therefore, it is essential to maintain human oversight and position generative AI as a supportive tool. Given the limited academic research and low integration of generative AI in existing data quality tools, this thesis provides insights and highlights the need for further research and responsible implementation in this area.

Keywords:
Artificial Intelligence; Data Quality; Data Quality Management; Generative Artificial Intelligence; Large Language Model

CERCS:
P170 Computer science, numerical analysis, systems, control; P175 Informatics, systems theory; P176 Artificial intelligence

Generatiivse tehisintellekti kasutamine andmekvaliteedi halduses

Lühikirjeldus:
Kvaliteetsed andmed on teadlike ja täpsete otsuste langetamiseks hädavajalikud, kuid kvaliteedi saavutamine ja säilitamine võib olla keeruline ja ressursimahukas protsess. Generatiivsel tehisintellektil on potentsiaal andmekvaliteedi halduse hõlbustamises ja käsitsi töö minimeerimises, võimaldades protsesse automatiseerida. Lõputöö uurib, kuidas saaks kasutada generatiivset tehisintellekti andmekvaliteedi halduse protsessides nende väljakutsete lahendamiseks. Töös esitletakse tüpoloogiat, mis töötati välja tuginedes turul saadaolevate andmekvaliteedi tööriistade süstemaatilisele analüüsile ja mida seejärel valideeriti ning täiustati kaasates andmekvaliteedi eksperte. Ainult 13 tööriista analüüsitud 88-st kasutavad praegu generatiivset tehisintellekti. Kuigi eksperdid nägid generatiivse tehisintellekti potentsiaali rutiinsete ülesannete automatiseerimises ja skaleeritavuse parandamises, väljendati kahtlusi usaldusväärsuse, selgitatavuse ja andmete turvalisuse pärast. Seetõttu on oluline säilitada inimesepoolne järelvalve ja kasutada generatiivset tehisintellekti toetava tööriistana, mitte iseseisva otsustajana. Arvestades teadusuurimuste piiratud arvu ja generatiivse tehisintellekti madalat kaasatust andmekvaliteedi tööriistadesse,

annab töö ülevaate generatiivse tehisintellekti kasutusvõimalustest ja rõhutab edasiste uurimuste ja vastutustundliku rakendamise vajadust.

Võtmesõnad:
Tehisintellekt; Andmekvaliteet; Andmekvaliteedi haldus; Generatiivne tehisintellekt; Suur keelemudel

CERCS:
P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria); P175 Informaatika, süsteemiteooria; P176 Tehisintellekt

Contents

1. Introduction

High-quality data is essential for reliable decision-making, operational efficiency, and strategic planning. However, recent industry surveys report that 77% of data professionals rate their data quality as "average at best" or worse, up from 66% the previous year [1]. Nearly half (49%) cite insufficient data management tools as one barrier to achieving high data quality [1]. Traditional data quality methods often rely heavily on manual processes, which are time-consuming, resource-intensive, and prone to human errors [2]. The emergence of generative artificial intelligence, particularly large language models (LLMs), offers new ways to automate data quality tasks, potentially making issue resolution more intuitive and accessible, and reducing the need for in-depth technical expertise [3, 4]. This shift is expected to allow experts to focus on more strategic activities, improving overall efficiency and effectiveness [2].

However, academic literature on the application of generative AI within data quality management remains scarce. While GenAI shows promise in various data-related tasks, its specific role in improving data quality management processes has not yet been systematically explored. This thesis addresses this gap by exploring how GenAI is currently used in data quality management, what its potential roles are, and what risks its adopters should be aware of. To support this objective, the study aims to answer the following research questions:

RQ1. How and to what extent is generative artificial intelligence currently integrated into data quality tools?

RQ2. What are the opportunities and risks of using generative artificial intelligence in data quality management?

To answer these questions, the study first systematically analyses existing data quality tools to assess how GenAI is currently used. Based on this, a typology of GenAI usage profiles is developed. This typology is then validated and refined through expert interviews, which also provide insights into the benefits and risks of adopting generative AI in data quality management practices. By identifying both opportunities and risks, this study aims to provide a balanced perspective on the role of GenAI in data quality management. In doing so, the study takes a step toward developing guidance for the responsible and informed adoption of generative AI technologies within data quality management processes. The remainder of this paper is organised as follows. Section 2 provides a theoretical overview, introducing foundational concepts related to data quality and generative AI. Section 3 outlines the methodology, detailing how the systematic analysis, typology and interviews were conducted. Section 4 presents the findings of the systematic analysis, the resulting typology, and insights from the expert interviews. Finally, Section 5 discusses the results in relation to the research questions and outlines the study's limitations.[1]

[1] ChatGPT was partially used throughout the thesis to improve wording and correct grammar.

2. Background and Related Work

This chapter provides the background for the study. The first subsection gives an overview of the key concepts, and the second outlines existing research relevant to this topic.

2.1 Theoretical Background

2.1.1 Data quality

Data quality is a multidimensional and complex concept that lacks a universally accepted definition. This ambiguity arises because data is used in diverse contexts, each with its own requirements, standards and expectations [5, 6]. Consequently, perceptions of data quality often reflect the priorities and perspectives of specific domains or user groups [6].

A commonly cited definition describes data quality as “fitness for use”, indicating that the concept of data quality is relative and depends on the context and the scenario [7]. Similarly, Wang and Strong [8] describe high-quality data as “data that are fit for use by data consumers”, reinforcing the notion that quality must be evaluated against the needs and expectations of those who rely on the data for decision-making or operational tasks.

Data serves as an input for deriving valuable information, which can be used to make informed, data-driven decisions [9]. However, the effectiveness of such decisions is highly dependent on the quality of the underlying data. According to a report by Drexel LeBow and Precisely [1], 76% of organisations cite data-driven decision-making as their top priority. At the same time, 67% of data and analytics professionals report that they do not fully trust the data their organisation relies on. This trust gap reflects the impact of poor data quality, which can result in inaccurate or misleading analysis results, ultimately making it difficult to draw reliable conclusions. Financially, the cost is also significant: Gartner [10] estimates that poor data quality costs organisations an average of 12.9 million dollars annually.

The level of data quality can be determined either subjectively by data consumers or objectively, using formal data quality management frameworks that consider organisational goals and operational needs [7, 11]. Data quality management (DQM) is a set of strategies, methodologies, and processes employed to ensure that data is accurate, reliable, and fit for decision-making, while meeting specific business requirements [12, 13]. Achieving perfect data quality across an entire organisation is neither economically feasible nor practically achievable, and in most cases, it is not even necessary [7, 14]. Therefore, rather than striving for perfection, organisations should aim for an optimal level of quality that aligns with their specific goals and use cases, and where the costs of further improvements are justified by the benefits [14, 15, 16].

To effectively manage data quality at an optimal level, one common approach is to evaluate data quality through specific dimensions, helping to prioritise improvements where they

matter most. Data quality dimensions are measurable data characteristics used to assess the overall quality of the data [11]. As data quality is a multidimensional concept, there is no universally accepted set of dimensions, and their definitions vary across different sources [17, 18]. Although the priority and implementation of data quality dimensions vary across domains, several core dimensions are generally considered universally relevant [11, 17, 18]. These include:

- accuracy, which refers to the extent to which data is correct, error-free and consistent with the real world [18].
- consistency, which refers to the uniformity of data representation across systems or contexts [19].
- completeness, which refers to the extent to which all required fields or values are present in the dataset [19].
- timeliness, which refers to the extent to which data is up-to-date and meets current requirements [19].

Traditionally, data quality has been managed through manual processes, requiring significant human effort and time for tasks such as cleansing, validation and monitoring. As the volume and complexity of data continue to grow, these rule-based approaches face limitations in scalability, adaptability, and cost-effectiveness [20, 21]. In response, (semi-)automated solutions, often incorporating machine learning and AI, have been integrated into DQM practices to enhance and streamline various data quality tasks [20, 22]. The emergence of generative AI has recently opened up new opportunities for transforming how data quality is assessed, corrected, and maintained.

2.1.2 Generative Artificial Intelligence

Generative Artificial Intelligence (GenAI) is a class of artificial intelligence models capable of producing new content, such as text, images and other content, based on underlying patterns learned from large datasets [23]. Unlike traditional AI models, which are primarily designed to analyse data and make predictions, GenAI systems are designed to generate human-like outputs. As generative AI technologies become increasingly accessible, they are starting to integrate into various sectors, including marketing, healthcare, education and entertainment [24, 25].

One of the most prominent applications of generative AI is Large Language Models (LLMs), which focus on text generation and, increasingly, visual content. Built on transformer-based neural networks, LLMs are trained on vast datasets containing data from various sources [25, 26]. These models generate coherent sequences of text by predicting the next word or phrase based on the context of preceding tokenised words [27]. These models learn language patterns through exposure to extensive examples and continuously improve, becoming more accurate and context-aware over time [25].

Although GenAI can assist in complex tasks, it also comes with risks related to ethics, technology, and regulations [28, 29]. One of the main problems is that generative AI is not fully reliable. This includes occurrences known as AI hallucinations, where the model

generates context that seems to be correct and believable, but is actually inaccurate or completely fabricated [30]. Such errors occur because generative models do not truly understand the topic [31] – they rely solely on patterns found in their training data, which may be biased, incomplete or outdated [25, 32]. Therefore, when generative AI is used, for example, in generating data quality rules or recommending transformations, humans must carefully review its output before implementation.

Despite the growing enthusiasm around using GenAI, its application in data quality management remains relatively underexplored, with only a handful of studies beginning to investigate this integration, such as [2, 3, 33]. Most existing studies lack a systematic approach or validation through real-world application and expert feedback. This study addresses that gap by combining a systematic analysis of market-available tools and discussions with data quality experts.

2.2 Related Work

In recent years, the potential of generative AI to support data-related tasks has received increasing attention. Within the context of data quality management, a growing number of publications have begun to explore possible applications of generative AI. These studies are largely conceptual or exploratory in nature. For example, Azeroual [2] provides a theoretical overview of GPT-4's possible benefits, risks, and use cases for generative AI, including error identification, data validation, and enrichment. Similarly, Varma et al. [33] present an enterprise-level application of LLMs in data management, where they are used for source identification, data profiling, rule generation, and SQL automation. Other works focus on domain-specific use cases. For instance, Krishnamoorthy [3] discusses the use of GenAI in healthcare, highlighting its role in automated data cleansing, standardisation, and the real-time integration of patient data from disparate sources. Additionally, Liu and Wongsosaputro [34] describe how GenAI supports regulatory reporting in the financial sector.

Yet, despite their contributions, these studies are generally limited to speculative scenarios and scattered use cases. They lack a comprehensive and real-world perspective on how generative AI is currently being adopted within DQM tools and practices. As a result, the field lacks a clear understanding of how GenAI functionalities are actually being implemented or utilised in practice.

In parallel, several studies have systematically analysed data quality tools. Ehrlinger and Wöß [35] conducted one of the first large-scale reviews, screening 667 tools and analysing 13 in detail. Martinsaari [36] focused on data warehouse contexts, identifying tools capable of automatic rule generation. Zhou et al. [37] examined data quality tools in machine learning workflows, identifying standard functionalities such as profiling, integration, and transformation. While these studies provide detailed overviews of available tools, they focus primarily on classical machine learning and AI automation, not on generative AI.

To the best of the author's knowledge, there is currently no survey or systematic review that focuses explicitly on the integration of generative AI within data quality tools. Nor has any existing work proposed a framework or typology to classify how GenAI could be used in data quality processes. This thesis seeks to address that gap by systematically analysing data quality tools on the market and developing a typology of GenAI usage profiles. The typology is further validated through expert input and complemented by an exploration of the associated opportunities and risks. By grounding the analysis in real-world tools and current implementations, the study offers a practical perspective on how generative AI is beginning to shape data quality management.

3. Methodology

This chapter presents the methodology used in the study, which consists of: (1) a systematic review of data quality tools, (2) the development of a typology of GenAI roles in data quality management based on the review, and (3) interviews with data quality professionals to validate the typology and assess broader potential and risks of GenAI in DQM. The following sections describe how the review, typology development, and interviews were planned and conducted.

3.1 Systematic Review of Data Quality Tools

The systematic review aims to provide an overview of existing data quality tools and to identify and examine those that leverage generative artificial intelligence. The systematic review is based on Kitchenham and Charters' guidelines [38]. The review addresses the following question:
RQ1: Whether and how GenAI is currently integrated into data quality tools?

3.1.1 Systematic Literature Review

Two complementary sources were used to build a list of data quality tools: academic publications and reputable online sources.

The academic review focused on literature retrieved from Web of Science and Scopus, as recommended by Brereton et al. (2007), Kitchenham and Charters (2007), and Carrera-Rivera et al. (2022), who emphasise their comprehensive coverage of computer science research and frequent use in systematic literature reviews. Search terms and the inclusion and exclusion criteria were defined prior to the review. The search terms used were kept as consistent as possible across databases, with adjustments made based on the characteristics and constraints of each platform. In Web of Science, the query was conducted using the terms "data quality" and "tool". In Scopus, the use of "data quality" and "tool" initially returned a very high volume of unrelated results. Therefore, a more specific phrase, "data quality tool", was used. In addition, a refined version of the "data quality" and "tool" query was applied, restricting the matches to titles, abstracts and keywords (see Table 1).

Table 1. Search queries used in Web of Science and Scopus.

Database	Query
Web of Science	"data quality" AND "tool"
Scopus	"data quality tool"
	"data quality" AND "tool"

The search was restricted to publications in English and written between 2015 and 2025. This approach was taken due to the relatively recent emergence of GenAI, making it unlikely that older tools would incorporate such technologies. To improve the relevance of the results and narrow the scope to the most relevant and applicable studies, results were filtered by subject domain. In Scopus, the search was limited to the computer science subject area. In Web of Science, only records within computer science-related categories were included. These categories were: *Computer Science Information Systems*, *Computer Science Theory & Methods*, *Computer Science Artificial Intelligence*, *Computer Science Interdisciplinary Applications*, *Computer Science Software Engineering*, *Computer Science Hardware Architecture*, and *Computer Science Cybernetics*.

The study selection process was based on the approach proposed by Brereton et al. [39], beginning with an initial screening, where papers were filtered based on their title and keywords. If a paper appeared relevant, its abstract and conclusion were reviewed to decide whether it fit the research scope. Finally, only documents that passed the abstract review were examined in full to confirm their relevance. Publications that explicitly mentioned data quality tools were included in the final selections. Metadata was then extracted for each of these studies and documented in a spreadsheet. Table 2 presents the metadata collected about the studies.

Table 2. Metadata collected from academic publications.

Metadata	**Description**
Name of the publication	Full title of the academic publication
Authors	Author(s) of the publication
Year of publication	The year the study was published
Mentioned tools	List of data quality tools referenced in the source
Found through (database)	The database where the study was found
Found through (search query)	Exact search query used to retrieve the source

In parallel, a targeted Google search was conducted to identify tools that were not mentioned in academic publications, but still relevant for this study. The search was carried out using the query "*top data quality tools in (2022 OR 2023 OR 2024 OR 2025)*". The sources were chosen based on their credibility, such as research platforms and well-known technology news outlets, and their relevance to the search query. Articles were included if they discussed multiple tools and provided descriptions or comparisons, rather than promotional content for a single product. To support data extraction, a spreadsheet was used to record metadata for each selected source. Table 3 presents the metadata collected from these sources. This approach ensured a broader coverage of existing tools, including those that may be widely used in practice but underrepresented in academic literature.

Table 3. Metadata collected from online sources.

Metadata	Description
Name of the article	Full title of the article
Website URL	Link to the source website
Year of publication	The year the article was published or last updated
Perspective	Type and intent of the article
Mentioned tools	List of data quality tools referenced in the source

3.1.2 Initial Screening

An initial analysis was conducted based on the information available on the tools' websites. Exclusion criteria defined to guide the selection of tools included: (1) tools with documentation available only in languages other than English, (2) tools with limited information available about the purpose or functionalities of the tool, (3) tools that have not been updated in the last two years, (4) tools that are not explicitly designed for data quality management or do not offer data quality functionalities (e.g., data integration or data management tools, data-debugging frameworks), (5) tools lacking executable components. The following data was recorded for each tool: the tool name, the company name, a link to the official website, the version, the release date, and the language.

3.1.3 In-Depth Documentation Review

The tools that passed the initial screening were subsequently reviewed and analysed based on documentation, websites, and product tours. Where available, free trials were also used to explore the tools' capabilities firsthand. The assessment criteria were informed by a synthesis of earlier systematic analyses of data quality tools [35, 36], with additional aspects, such as technical capabilities, introduced by the author. These criteria fall into four main categories: core data quality functions (e.g., cleansing, standardisation), monitoring and profiling capabilities, technical characteristics (e.g., supported data formats or sources), and user experience features (e.g., visualisation, custom rules, and virtual assistance). Each tool was assessed using category-specific values. For most criteria, the following scale was applied:

- yes: functionality was present;
- yes (with GenAI): functionality was enhanced by GenAI;
- no: functionality was absent.

The complete list of evaluation criteria is provided in Table 4.

Table 4. Evaluation Criteria for the Integration of Generative AI in Data Quality Tools.

Functionality	Description	Value
(1) Core data quality features		
Data cleansing	Data cleansing is the process of fixing or removing incomplete, incorrect, inaccurately formatted, or duplicated data [40].	Yes; Yes (with GenAI); No
Data standardisation	Data standardisation is the process of converting data into a common format, ensuring consistency across different datasets [41].	Yes; Yes (with GenAI); No
Data integration	Data integration is the process of combining data from different sources into a unified and consistent format [42].	Yes; Yes (with GenAI); No
Supported data quality dimensions	This refers to the data quality dimensions the tool supports (e.g., accuracy, completeness, consistency, uniqueness, timeliness).	List of supported dimensions; -
Rule-based checks	Rule-based checks are predefined rules or conditions applied to data to validate its accuracy, consistency and compliance with business requirements or data quality standards [43].	Yes; No
(2) Data quality monitoring		
Anomaly detection	Anomaly detection is the process of identifying unusual patterns or outliers in data that do not conform to usual or expected behaviour [44].	Yes; Yes (with GenAI); No
(Real-time) Data quality monitoring	Data quality monitoring refers to the ongoing process of tracking and evaluating the quality of data within an organisation to ensure it meets established standards and business requirements [45].	Yes; Yes (with GenAI); No
Data profiling	Data profiling is the process of examining data to collect statistics and identify characteristics within a dataset [6].	Yes; Yes (with GenAI); No
(3) Technical capabilities		
Supported data formats	This refers to the data formats the tool can handle (e.g., CSV, JSON, XML).	Various; List of formats if only a few are supported.
Supported data sources	This refers to the data sources that can be connected or integrated with the tool (e.g., databases, APIs, cloud platforms).	Various; List of sources if only a few are supported.
Scalability	Scalability indicates whether the tool can handle large volumes of data efficiently.	Yes; No
(4) User experience		

Functionality	Description	Value
Data visualisation	Data visualisation refers to the ability to present data quality results through charts, dashboards, and reports.	Yes; Yes (with GenAI); No
Virtual assistance	Virtual assistance includes chatbot or copilot features that support data quality tasks through natural language interfaces.	Yes; Yes (with GenAI); No
Ability to create custom quality rules.	This refers to the user's ability to define and implement custom data quality rules within the tool.	Yes; Yes (with GenAI); No

3.2 Typology

The development of the typology in this study was inspired by the ideal-type analysis approach described by Stapley et al. [47]. The process began with familiarisation with the data, followed by type construction, validation, and refinement. Additionally, the structure of the profiles (or types) was guided by the conceptual framework for AI-based ideation systems proposed by Lehmann et al. [48], in which each role is characterised by the pain it addresses, the benefits it offers, and the risks it entails.

In practice, the typology was developed using findings from the systematic analysis of data quality tools, with a specific focus on the integration of GenAI within these tools. The aim was to categorise GenAI based on its roles and contributions to enhancing data quality. The process began with a detailed review of findings from the analysis of data quality tools. Key findings were synthesised into profiles designed to reflect different ways in which GenAI can contribute to data quality processes. Each profile included a description of the pain point it addresses and the role of GenAI as a pain reliever.

To ensure the validity of the typology, interviews were conducted with data quality professionals. The experts provided feedback on the relevance and applicability of the profiles and were also asked to identify potential benefits and risks associated with each one. Based on this feedback, the typology was refined to better capture the practical use cases and challenges of using GenAI in the context of data quality.

3.3 Interviews

To complement the systematic analysis, semi-structured interviews were conducted with data quality professionals to gain insights into the advantages and disadvantages of using generative AI in data quality analysis. This approach was employed to explore the potential and challenges of integrating generative AI into data quality management.

Participants were selected using purposive sampling, targeting individuals with expertise in data quality. The sample included both data quality practitioners and academic researchers with practical experience in the field. To ensure a comprehensive and diverse range of

perspectives, the sample targeted individuals from various roles and different application domains. Geographic diversity was also considered essential: rather than focusing on one country, participants were drawn from different regions to ensure global relevance. Additionally, the sample included experts with varying levels of experience to capture both fresh viewpoints and long-standing expertise. Interviewees were identified based on their publicly available profiles on networking platforms and their authorship of publications related to data quality. The interviews were conducted remotely via Microsoft Teams in April 2025.

Given the richness of the data gathered and the diversity of the sample, the number of interviews was considered methodologically sufficient. Prior research has shown that small, targeted samples can be appropriate in qualitative research when participants are purposefully selected for their in-depth knowledge [49, 50]. Guest, Bunce, and Johnson [49] found that thematic saturation, the point at which no new themes emerge, can be reached within the first 6 to 12 interviews. Similarly, Malterud, Siersma, and Guassora [50] emphasise the concept of "information power", which suggests that the more relevant and substantial the information provided by participants, the fewer individuals are required to obtain meaningful results. Moreover, Mason [51] notes that Ph.D.-level qualitative studies commonly justify sample sizes between 5 and 30 interviews, depending on the study's scope and depth. In this context, the interviews conducted for this study provided sufficient insights to identify common themes and challenges related to data quality management.

The interview guide focused on four key areas:

1. the current state and challenges of data quality management: participants were asked about the common obstacles they face in ensuring data quality;
2. the potential of generative artificial intelligence to improve data quality management: questions were aimed at exploring participants' perceptions of using GenAI in data quality processes;
3. the validation of typology: experts were asked to evaluate whether the developed profiles for GenAI were realistic and applicable to current or future uses of GenAI in the field;
4. the associated risks and limitations of using GenAI: questions exploring the possible drawbacks and challenges associated with integrating GenAI into data quality processes.

The interview questions were informed by insights from the literature on data quality management and the preliminary findings of the systematic tool analysis. The interview questions are available in Appendix A. The interviews were recorded and transcribed prior to analysing the data. All participants were informed about the purpose of the study and their rights, and they provided informed consent prior to the interviews. The automatic transcription provided by Teams was used and then manually reviewed to ensure it had been transcribed accurately. The transcriptions were then returned to the participants for verification, allowing them to clarify their responses. The transcriptions were then analysed using thematic analysis, following the approach outlined by Braun and Clarke [52], which

supports identifying, analysing, and reporting themes within qualitative data. Initial codes were developed through close reading, and similar codes were grouped into broader themes, which formed the basis for the results presented in the following chapter.

4. Results and Analysis

This chapter presents the results of the study, covering three main parts: (1) the findings from the systematic review of data quality tools, (2) the GenAI role typology developed based on this review, and (3) insights from expert interviews, which served to validate the typology and provide additional perspectives on the opportunities and risks of GenAI in DQM. The following sections present and analyse the outcomes of each part in detail.

4.1 Systematic Analysis of Data Quality Tools

4.1.1 Literature Search

To gain a comprehensive understanding of the landscape of data quality tools, a systematic literature search was conducted using both academic databases and industry-oriented sources.

Two academic databases, Scopus and Web of Science, were used to identify relevant publications for this review. In *Web of Science*, the search query (SQ1: “DATA QUALITY” AND “TOOL”), limited to the computer science subject area, English language, and publications from 2015 to 2025, returned 387 results. In *Scopus*, an initial search was conducted using the query (“DATA QUALITY” AND “TOOL”) with subject area limited to computer science and the publication years 2015-2025, yielding over 15,000 documents. To narrow it down, a more specific query (SQ2: "DATA QUALITY TOOL") was tested, returning 386 publications. To align with the parameters used in Web of Science, a refined search query was constructed:

(SQ3: TITLE-ABS-KEY (“DATA QUALITY” AND TOOL) AND PUBYEAR > 2014 AND PUBYEAR < 2026 AND LIMIT-TO (SUBJAREA, “COMP”) AND LIMIT-TO (LANGUAGE, “English”) AND LIMIT-TO (EXACTKEYWORD, “Data Quality”)).

This refined search returned 653 documents.

All retrieved articles underwent a screening process. To avoid duplication, results from all three queries were cross-checked, and duplicate entries were removed. As shown in Figure 1, a total of 53 unique publications were ultimately identified as relevant for inclusion in this review. These publications referenced a total of 159 distinct data quality tools.

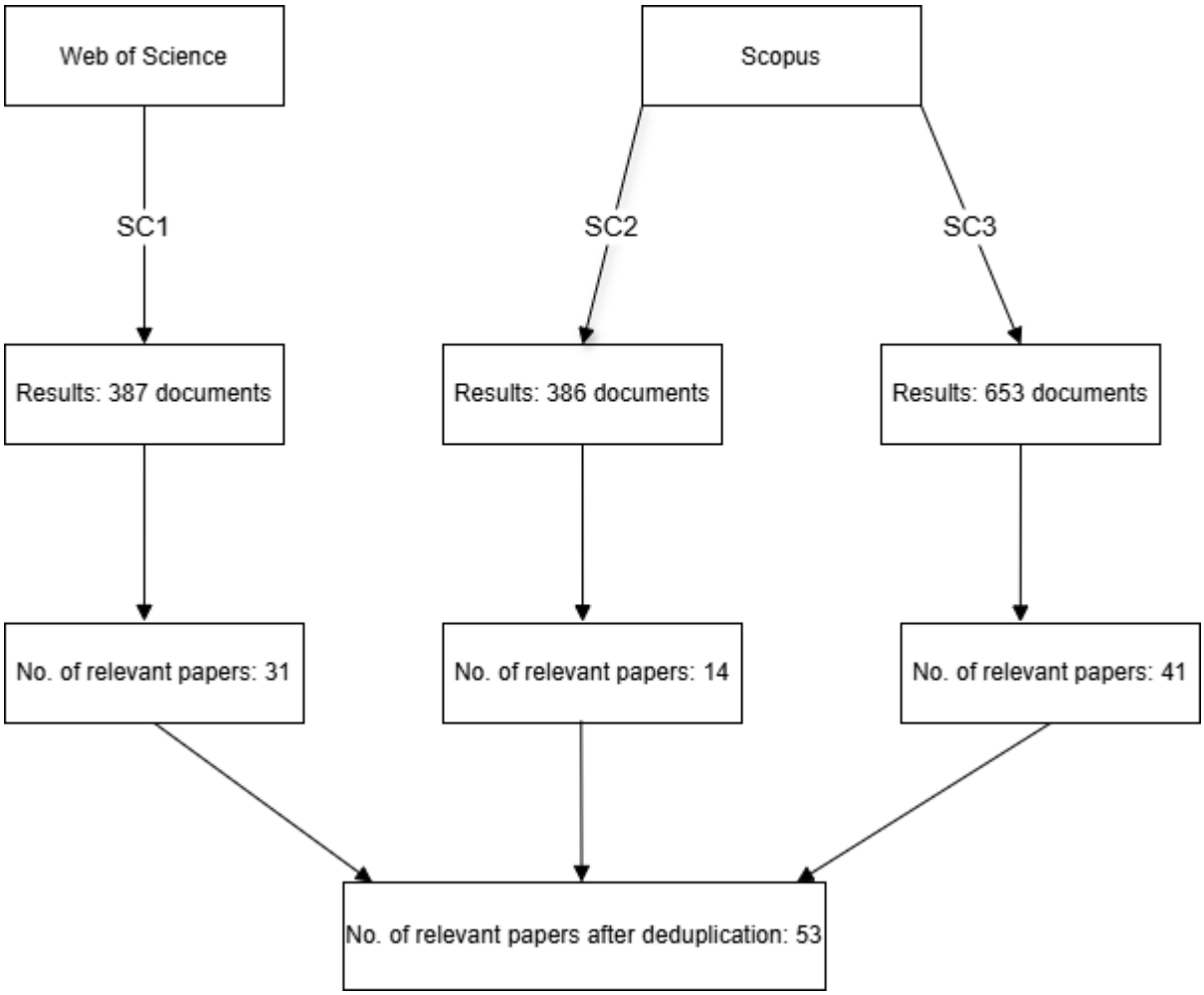

Figure 1. Overview of the systematic literature search.

A detailed list of the relevant literature is available in Appendix B.

To capture industry trends and tools not covered in academic publications, a Google search was conducted according to the design described in section 3. This search aimed to identify recent articles listing or reviewing data quality tools. The first ten pages of the Google search were reviewed to ensure coverage of the most relevant and current sources. Articles were included if they focused on data quality tools, introduced tools not already captured through academic sources, and were published by reputable technology news outlets, software review platforms, or research-oriented websites. This process yielded 17 relevant articles, discussing 122 unique tools. The inclusion of grey literature helped to supplement academic publications by reflecting currently popular data quality tools on the market. A complete list of the identified sources is provided in Appendix C.

4.1.2 Initial screening

The literature review identified 209 unique data quality tools. Each tool was assessed against the exclusion criteria defined in the methodology chapter. If a tool met one or more of the

exclusion criteria, it was removed from further consideration. The number of tools excluded under each criterion is summarised in Table 5.

Table 5. Exclusion criteria and the number of tools excluded.

Exclusion Criteria	Description	Number of tools excluded
EC1	Tools with documentation available only in languages other than English.	0
EC2	Tools with insufficient available information.	28
EC3	Tools that have not been updated within the last two years.	31
EC4	Tools that are not primarily designed for data quality purposes.	42
EC5	Tools that do not have any executables.	41

EC1 did not lead to exclusions, as all identified tools had documentation available in English.

EC2 resulted in the exclusion of 28 tools that were mentioned in the literature but had no accessible website or documentation. In some cases, websites existed but lacked sufficient information about functionalities, integrations, or usability. Many of these tools seemed not to exist anymore, possibly discontinued.

EC3 excluded 37 tools that had not been updated within the last two years. As generative AI is a relatively new concept in data quality, outdated tools were considered less relevant to this study.

EC4 accounted for 49 exclusions, as these tools were primarily designed for purposes outside the scope of traditional data quality management, such as data visualisation, knowledge graph assessment, data lineage, integration, or analytics. As a result, they fell outside the scope of this study, which focuses on analysing data quality tools.

EC5 led to the exclusion of 45 tools that were primarily research projects. While they were well-documented in academic papers, they were not fully developed and lacked available executables, making them unusable in the real world.

After applying all exclusion criteria, 88 tools out of 209 remained for documentation review.

4.1.3 Documentation review

A total of 88 tools were reviewed, with an initial focus on identifying whether references to the use of generative artificial intelligence were present in their documentation and websites. Following this screening process, 15 tools were selected for further analysis based on the

predefined criteria, as they already leverage generative AI or have announced plans to implement it in upcoming versions (see Table 6).

Table 6. Tools that use or claim to start using GenAI.

Tools that use generative AI	Website
Acceldata Data Observability Cloud	link
Alation Agentic Data Intelligence Platform	link
Ataccama ONE	link
Bigeye	link
Collibra Data Quality & Observability	link
HubSpot Operations Hub	link
Monte Carlo	link
QuerySurge	link
Soda	link
Dataedo	link
DQLabs Platform	link
Informatica Intelligent Data Management Cloud	link
Zoho DataPrep	link
Datactics	link
Aperture Data Studio	link

The opportunity to test the tools was limited. Of the 15 tools reviewed, only four, Soda, QuerySurge, HubSpot, and Zoho DataPrep, offered publicly accessible free trials. For others, it was possible to request free trials (Acceldata) or product demos, namely: Alation, Ataccama, Bigeye, Collibra, Monte Carlo, Dataedo, DQLabs, Informatica, Datactics, and Aperture. However, these requests were not fulfilled, likely due to the educational nature of the inquiry. Some limited access was available through interactive product tours offered by Collibra, Monte Carlo, and Datactics.

Following this, these tools were evaluated according to predefined criteria for their functionalities. Table 7 presents an overview of the generative AI functionalities identified across the reviewed data quality tools.

Table 7. GenAI functionalities across data quality tools.

Functionality	Number of tools	Tools
Quality rule generation	8	Acceldata Data Observability Cloud, Ataccama ONE, Collibra Data Quality & Observability, HubSpot Operations Hub, Monte Carlo, Soda, DQLabs Platform, Informatica Intelligent Data Management Cloud
Virtual assistance	6	Ataccama ONE, HubSpot Operations Hub, Soda, DQLabs Platform, Informatica Intelligent Data Management Cloud, Zoho DataPrep
Data standardisation	3	Ataccama ONE, Informatica Intelligent Data Management Cloud, Zoho DataPrep
Data profiling	2	Monte Carlo, Informatica Intelligent Data Management Cloud
Data integration	1	Informatica Intelligent Data Management Cloud
Data cleansing	1	Zoho DataPrep

Based on the analysis, generative artificial intelligence is most commonly used for **generating quality rules**, with this functionality present in **8 tools**. In these cases, users define data quality requirements in natural language, which GenAI then translates into executable data quality rules. The second most frequent use case is **virtual assistance**, identified in **6 tools**, enabling users to ask data-related questions, identify the root causes of issues, understand how to resolve them, gain insights into how to use the tool, and interact with the documentation. Other GenAI-enabled functionalities identified include:

- **data standardisation**: e.g., asking GenAI to transform data into a specific format;
- **data profiling**: such as requesting summary statistics or identifying outliers;
- **data integration**: e.g., generating integration mappings;
- **data cleansing**: automating error detection and correction tasks.

In addition to the functionalities outlined in the predefined criteria, it was found that generative AI is also utilised in tools for **generating metadata**, **asset descriptions**, and **queries**, as well as for **writing validation tests**, **generating synthetic datasets** and **sample data**.

Furthermore, aside from the tools currently claiming to leverage GenAI, Datactics and Aperture Data Studio reported that GenAI products are under development. Datactics is working on SQL rule creation based on natural language input, and Aperture Data Studio is developing GenAI for rule creation, profiling insights, and metadata autofill. Collibra has also announced a forthcoming GenAI copilot.

Table 8 shows an overview of GenAI integration for each tool, showing which specific functionalities are supported. A checkmark (✓) indicates GenAI-supported functionality.

Table 8. Tools using GenAI and their supported functionalities.

TOOL NAME	Data cleansing	Data standardisation	Data integration	Rule-based checks	Anomaly detection	(Real-time) DQ monitoring	Data profiling	Scalability	Visualisation	Virtual assistance	Alerts/notifications	Ability to create custom DQ rules	Notes
Acceldata Data Observability Cloud												✓	GenAI for metadata generation.
Alation Agentic Data Intelligence Platform													GenAI for asset descriptions and title generation.
Ataccama ONE		✓								✓		✓	GenAI for test data, asset descriptions and query generation.
Bigeye													GenAI for improvement suggestions and issue descriptions.
Collibra Data Quality & Observability												✓	GenAI for asset descriptions. GenAI copilot is in development.
Datactics													GenAI for SQL rule creation is in development.
HubSpot Operations Hub										✓		✓	GenAI for task automation and workflow generation.
Monte Carlo							✓					✓	GenAI for recommending monitors based on profile, query history and metadata.
QuerySurge													GenAI for data validation tests.
Soda										✓		✓	
Aperture Data Studio													GenAI for rule creation, profiling insights and metadata autofill is in development.
Dataedo													LLMs for generating documentation.

TOOL NAME	Data cleansing	Data standardisation	Data integration	Rule-based checks	Anomaly detection	(Real-time) DQ monitoring	Data profiling	Scalability	Visualisation	Virtual assistance	Alerts/notifications	Ability to create custom DQ rules	Notes
DQLabs Platform										✓		✓	
Informatica Intelligent Data Management Cloud		✓	✓				✓			✓		✓	
Zoho DataPrep	✓	✓								✓			GenAI for generating synthetic datasets, finding external data.

4.1.4 Typology

Based on the most frequently observed generative AI functionalities in current data quality tools, identified through the systematic analysis, four usage profiles were developed: Translator, Explainer, Resolver, and Integrator (see Table 9). These profiles reflect the use cases for GenAI in data quality contexts.

Table 9. Usage profiles of GenAI.

	Profile	Description
1.	GenAI as a translator	GenAI converts natural language inputs into machine-readable data quality logic, including data quality rules, validation tests, and profiling queries. *Observed in: Acceldata Data Observability Cloud, Ataccama ONE, Collibra Data Quality & Observability, HubSpot Operations Hub, Monte Carlo, Soda, DQLabs Platform, Informatica Intelligent Data Management Cloud, Zoho DataPrep.*
2.	GenAI as an explainer	GenAI provides real-time, natural language explanations to data quality analysis results, including potential root causes. *Observed in: Ataccama ONE, HubSpot Operations Hub, Soda, Informatica Intelligent Data Management Cloud, Zoho DataPrep.*
3.	GenAI as a resolver	GenAI automates the identification and resolution of data quality issues. *Observed in: Zoho DataPrep.*
4.	GenAI as data integrator	GenAI aligns and integrates data across different

	Profile	Description
		systems and schemas. *Observed in: Informatica Intelligent Data Management Cloud.*

To assess the practical relevance of these profiles, they were introduced to interview participants during the interviews. Participants were asked to reflect on the realistic nature, usefulness, and potential risks of each profile. Their feedback offered valuable insights into how these conceptual profiles correspond to real-world scenarios. Based on this input, the refined and validated typology is presented in Section 4.3.

4.2 Interview Results

A total of seven experts were interviewed, representing a range of backgrounds and experiences. The sample included research associates (working on data-related topics but not necessarily focused on data quality), data managers, and R&D engineers, with professional experience in data quality management varying from over 20 years to participants currently pursuing their PhDs. This diversity of perspectives enriched the findings, reflecting both hands-on practitioners and those with a more academic focus. The participants also reflected geographic diversity, with experts based in Northern Europe, Central Europe, and the United States. The general profiles of the experts are shown in Table 10. Despite the relatively small sample size, the participants provided rich insights, and no substantially new themes emerged in the later stages of the interviews.

The interviews were conducted in April 2025 via Microsoft Teams, with each session lasting between 35 and 45 minutes. The responses have been categorised into three main themes: the potential of generative AI in data quality management, associated risks and downsides, and the reflections on the typology developed in the study. These themes were intentionally addressed during the interviews, as participants were asked to comment on those aspects.

Table 10. Interview participants' general profiles.

Respondent ID	Job position	Academic qualification	Years of experience	Experience with GenAI in DQM
R1	PhD Student	Master's degree, pursuing a PhD.	1 year	Yes
R2	R&D Engineer	Dr.	10+ years	No
R3	Research associate	Master's degree, pursuing a PhD.	1.5+ years	No
R4	Researcher	Dr.	20+ years	No
R5	Research associate	Master's degree, pursuing PhD.	3-4 years	No

Respondent ID	Job position	Academic qualification	Years of experience	Experience with GenAI in DQM
R6	Data manager	Master's degree, pursuing PhD.	4 years	Yes
R7	Researcher	Dr.	10 years	Yes

Three of the seven respondents reported experience with generative AI in data quality management. One respondent had used GenAI to generate SQL queries based on synthetic data, while others were involved in projects focused on leveraging LLMs for data evaluation, error detection, and improvement suggestions.

4.2.1 The Potential of Generative AI in Data Quality Management

Interview respondents highlighted several benefits of using generative AI in the context of data quality management. Table 11 provides an overview of identified benefits and the respondents who mentioned them.

Table 11. Key benefits of GenAI in data quality management.

Identified benefit	Respondents
Automation of manual data quality tasks	R1, R2, R3, R4, R5, R6, R7
Improved scalability	R1, R3, R4, R5, R6
Data enrichment	R3, R5, R6, R7
Ability to understand context and identify patterns.	R5, R6, R7
Higher consistency and reliability for repetitive tasks	R1, R3, R6
Support for unstructured and multimodal data	R3, R5
Increased annotation accuracy	R1
Cost-reduction	R1
Ability to generate synthetic data to avoid privacy issues	R4
Ability to gather insights about data	R2

The most frequently mentioned benefits were the **automation of manual and low-ambiguity tasks**, as well as **enhanced scalability**. As data volumes grow, human capacity to manage and validate data becomes increasingly limited. In this context, automation is a key enabler of scalability, allowing systems to handle larger datasets without a proportional increase in manual effort. Respondent R3 noted that despite the increasing digitisation, maintaining and ensuring data quality is still largely manual, making it resource-intensive, costly, and error-prone, even with expert involvement. Using GenAI in these contexts would

allow organisations to automate repetitive tasks, such as filling in missing modalities or removing duplicate values. This could reduce the burden on data stewards and enable domain experts to focus on higher-value activities.

Another mentioned benefit was GenAI's ability to **identify patterns** in data. This is especially useful in complex, domain-specific datasets. For instance, R6 described a situation in healthcare where a patient's record contained conflicting information: the patient was marked as obese, even though other data fields, such as weight and fitness level, completely contradicted this classification. This type of data issue is difficult to detect and fix using traditional rule-based checks, as it requires contextual understanding.

GenAI is not limited to structured data. When asked about the current challenges in data quality management, interviewees R3 and R5 mentioned the increasing presence of unstructured and multimodal data. Each modality requires distinct handling. GenAI was seen as a promising solution to address this complexity, as it could help **integrate diverse data formats**, such as videos, presentations, PDFs, or process data, and transform unstructured sources into structured formats. It could also enable multimodal semantic search, allowing users to query and utilise data across data types, regardless of the original format (R5).

In addition to capabilities directly related to data quality improvement, several respondents noted that GenAI could also support data enrichment, particularly through metadata generation and contextual annotation (R3, R5, R6, R7). For example, R5 highlighted that GenAI could assist in metadata extraction and enrich data with contextual information, thereby reducing manual effort and increasing efficiency. Similarly, R3, R6, and R7 emphasised GenAI's potential to generate metadata based on domain-specific context and business rules.

Other advantages included **increased annotation accuracy** (R1), **cost-reduction** (R1), the ability to **generate synthetic data** to address privacy concerns (R4), and **data insight generation** (R2).

4.2.2 Risks and Downsides of Generative AI in Data Quality Management

While generative AI introduces significant potential for improving data quality processes, the study also investigated its associated risks. Table 12 provides an overview of identified limitations and the respondents who mentioned them.

Table 12. Key risks of GenAI in data quality management.

Identified risk	Respondents
Lack of trust in outputs	R1, R2, R3, R4, R5, R6, R7
Security and privacy concerns	R1, R2, R4, R5, R6, R7
Legal risks, compliance	R1, R5, R6, R7

Identified risk	Respondents
Cost	R1, R4, R6, R7
Over-reliance on GenAI tools	R1, R3, R4, R6
Training data quality concerns	R1, R4, R5, R6
Lack of explainability	R1, R4, R5
Ownership and responsibility issues	R5

One of the most prominent concerns is the **lack of trust** in the outputs produced by generative AI models. All respondents mentioned the risk of inaccurate outputs and model hallucinations. The black-box nature of these models poses a challenge for **explainability** as well. This means that the internal mechanisms of generative AI systems are often not transparent, making it difficult for users to understand why and how specific outputs are generated (R4).

A second major category of risks relates to **privacy**, **security**, and **compliance**. Integrating generative AI into workflows involving personal or sensitive information brings legal and ethical challenges. As R5 pointed out, many organisations are concerned about using GenAI because of potential non-compliance with data protection regulations, such as the General Data Protection Regulation (GDPR).

Since many GenAI tools rely on third-party APIs, organisations may accidentally expose confidential and sensitive data. Therefore, as most respondents (R1, R2, R4, R5, R6) noted, these models should be used with great caution, or not used at all, to process sensitive data, like patient records. However, controlling employee usage of public GenAI tools can be difficult.

To minimise these risks, one potential solution is to deploy on-premise or private generative AI models. This way, the risk of leaking sensitive information can be reduced, as the model would be tailored for specific tasks or datasets, preventing unintentional data exposure. However, R1 and R6 raised concerns that deploying such solutions comes with high infrastructure, model training and maintenance **costs**, which may reduce their accessibility or appeal for certain organisations.

A generative AI model's ability to produce trustworthy outputs depends heavily on the quality of the data it was trained on. Concerns such as **data pollution**, **data poisoning**, and **bias** in training datasets were cited as threats to both the accuracy and fairness of outputs (R1, R5, R6).

As R4 said, "There's still a lot of research needed to make generative AI more sustainable".

4.3 GenAI x Data Quality Management Typology

Based on the results of the systematic analysis, four ideal-type profiles for the application of generative AI were developed: Translator, Explainer, Resolver, and Integrator. These profiles were subsequently presented to data quality experts for feedback.

Overall, the profiles were generally well received, with participants recognising their relevance and usefulness in data quality management practices. No significant structural changes were proposed, however, some participants found certain descriptions unclear and suggested refinements to define the scope of each role better. For instance, in the Translator profile, there was discussion around whether the role should be limited to translating natural language into data quality rules or whether it should more broadly encompass the conversion of natural language into executable programming logic, including tasks such as standardisation, formatting, and validation. Among the four profiles, Translator, Explainer, and Resolver were perceived as the most promising. The Integrator profile received a more cautious response: while seen as potentially useful in some scenarios, it was generally regarded as less transformative than the other roles. To provide an overview, Table 13 summarises the revised profiles, including their descriptions, common use cases, and risks based on expert feedback.

Table 13. Summary of Generative AI Roles in data quality management.

Role	Description	Key Use Cases	Main Risks
Translator	Converts natural language inputs into actionable programming steps.	Creating data quality rules, standardisation, and formatting.	Misinterpretation of inputs, inaccurate outputs, and unclear ownership.
Explainer	Translates technical outputs into accessible summaries.	Explaining outliers, describing data distributions, providing root-cause analysis, and aiding non-technical users.	Oversimplification, unverifiable explanations, and user overconfidence.
Resolver	Suggests or applies data quality fixes based on learned patterns.	Deduplication, inconsistent units, and handling missing data.	Inaccurate suggestions, context-insensitive fixes, and limited traceability.

Role	Description	Key Use Cases	Main Risks
Integrator	Maps and aligns data across sources.	Schema matching, field mapping, and prompt-based integration guidance.	False mappings, low added value in mature setups, and automation is not always necessary.

The following sections provide a detailed examination of each role, outlining typical use cases and key benefits and risks.

4.3.1 GenAI as a Translator

In this role, generative artificial intelligence acts as a bridge between the user and the technical system. The technical complexity of implementing data quality checks often limits participation to those with specific programming skills. Generative AI offers a potential solution by enabling users to describe their requirements in natural language, which the system can translate into executable rules. This includes tasks such as verifying the uniqueness of records, enforcing formatting standards, retrieving specific values, applying standardisation or other quality rules, or creating validation tests.

Interviewees highlighted several potential benefits of using generative AI for data quality management. One key advantage is increased accessibility: domain experts who may lack the technical skills to write rules can still contribute by describing their needs in natural language. That way, GenAI bridges communication between stakeholders and technical teams. Another mentioned benefit is increased efficiency and speed. This is especially useful in helping teams keep pace with ever-evolving business requirements.

However, this approach introduces certain risks. One concern is the overreliance on the AI's contextual understanding, which can lead to incorrect or imprecise outputs as the model may not take every relevant factor into account. Another issue is the potential for overly broad outputs due to imprecise natural language input. These concerns point to the importance of keeping domain experts involved in the process, commonly referred to as a human-in-the-loop approach. Several participants also emphasised the necessity of expert validation. As R3 put it, "So it's important to still keep experts in the loop. I think these models can be really helpful if we take the rule creation example again, you cannot use all your experts to write DQ rules all day, they have better things to do. So, although it would be amazing if you could have an AI model to introduce new rules without any validation, it might also cause a lot of issues."

4.3.2 GenAI as an Explainer

As an explainer, generative artificial intelligence acts as an assistant that summarises distributions, anomalies, and key insights, and assists users in quickly identifying and retrieving the most relevant datasets or documents. This can be particularly valuable for non-technical users, who may lack the skills and time to navigate complex datasets.

By translating complex statistical outputs into simplified summaries, GenAI supports data literacy and increases transparency in data quality pipelines, helping users become more familiar with the issues. It democratises understanding by making technical insights more accessible to non-experts, eases root-cause analysis by offering explanations and saves time by providing quick entry points into problem-solving. As R5 noted,
"As an employer, you are responsible for something, but you don't, especially for the ideation part, you have no clue what happened, and you have a chatting partner. It's more like, even if they don't give me a result, I can have some starting point."

However, interviewees also mentioned several risks. One key concern is over-simplification, when translating information into natural language, some details may be lost, leading to misleading conclusions. Additionally, R4 worried that there is no way to validate GenAI's explanations. Another issue is overconfidence, where users may develop a false sense of expertise and start to see themselves as experts without fully understanding the underlying complexities. Because of that, GenAI should not be used as a substitute for expert interpretation, but rather as a support tool that helps users engage with data.

To ensure responsible use, R7 emphasised the importance of raising user awareness that GenAI is not perfect. While it can help users identify problems and generate hypotheses, it should not be the sole basis for decision-making.

4.3.3 GenAI as a Resolver

The third type refers to generative artificial intelligence acting as a resolver of data quality issues. Manually detecting and resolving data quality problems can be both time-consuming and error-prone. GenAI can support this process by suggesting context-aware solutions derived from patterns within the dataset. This includes auto-correcting inconsistent units or removing duplicated data.

Interviewees identified several benefits of this approach. Most importantly, the use of GenAI can reduce manual workloads, which frees up experts to focus on more complex tasks. It can also increase speed, consistency and scalability. Unlike human experts, whose approaches may vary over time, between individuals, or as teams and processes evolve, GenAI can help ensure consistent fixes are applied across the board, thereby supporting the long-term stability of data quality standards.

However, respondents also emphasised several downsides. One core issue is that GenAI, while context-aware to some extent, lacks the nuanced understanding that human experts bring, particularly in complex cases. As a result, its suggestions may not always be accurate, and if applied automatically, could worsen data quality rather than improve it.

R5 noted that GenAI is most effective in structured and predictable environments, where fixes can be derived from historical data. In more chaotic contexts, its performance may be unreliable. It was also mentioned that the greatest value lies not in automating fixes, but in suggesting ideas that experts can then evaluate, refine or reject (R6, R7). As R6 said, "Removing duplicate values is a pretty standard and straightforward part of data analysis - not a particularly complex task. Using generative AI for such simple issues may not bring much added value, since a small code snippet can do the job just fine. But AI could be useful when it's unclear what problems need to be solved, for example, by suggesting ideas for data cleaning. Based on those ideas, more precise rules and solutions can be defined manually."

There are also technical challenges to consider. These include traceability and integration limitations with older systems that do not support real-time GenAI integration (R3). Furthermore, model collapse may affect GenAI's ability to make accurate recommendations (R4).

In conclusion, while GenAI has strong potential as a resolver, it is best seen as a support tool for generating suggestions rather than as an autonomous fixer.

4.3.4 GenAI as an Integrator

In this role, generative artificial intelligence could assist in mapping and aligning data across different databases and systems. It's often a time-consuming task involving large volumes of data. GenAI can support data integration processes such as schema matching, attribute alignment and the generation of data integration mappings based on natural language prompts.

Interviewees indicated that this application is valuable in contexts where documentation is incomplete or inconsistent. By analysing not only attribute names, but also data content, metadata and other contextual clues, GenAI can help identify relationships that may not be immediately obvious, thereby enhancing data discoverability. As R3 explained, "Many companies, especially large, well-established ones, face the issue of having data siloed in different systems, and it can be very difficult to link them. … This is one of the main challenges that companies face: they're working with data in parallel but don't know that other parts of the organisation have valuable data they could be using. GenAI could be very helpful for identifying and linking such data across silos."

The benefits are similar to other use cases: reduced manual effort, improved scalability, and time savings. GenAI can approach integration more flexibly by incorporating broader

context and patterns. R7 noted that, while machine learning has long enhanced these capabilities, the introduction of language models provides a significant additional boost.

However, interviewees also cautioned against fully automating such pipelines. One significant risk is the possibility of incorrect mappings and false matches. Furthermore, not all organisations need this level of automation. R6 noted that many integration tasks are already well-defined and documented in larger enterprises. Human oversight remains important for foundational tasks to ensure users understand and trust their data pipelines.

Overall, while the initial typology was generally well received, the expert interviews enabled valuable refinements to improve clarity and better align the profiles with real-world data quality practices. Notably, no additional roles were proposed, but the Integrator profile was viewed as less impactful than the others.

5. Summary and Discussion

This chapter summarises and discusses the key findings in relation to the research questions and the existing literature. The discussion is structured around the key themes that emerged from both the systematic analysis and the interviews. Finally, the limitations of the study are outlined.

RQ1. How and to what extent is generative artificial intelligence currently integrated into data quality tools?

The systematic analysis of selected data quality tools revealed that GenAI integration in DQM is still in its early stages. The initial set of 209 candidate tools was identified through academic databases and credible online sources. The tools were evaluated against defined criteria. 121 tools were excluded for reasons such as being inactive research prototypes, lacking publicly available documentation, having no updates within the last two years, or not being relevant to data quality management. This left 88 tools for a more detailed analysis. Among the remaining 88 tools, only 13 demonstrated GenAI integration. Additionally, 2 indicated plans to implement GenAI features in the near future. This means that only around 15% of the analysed tools have incorporated GenAI into their functionality.

Among the analysed tools, the main GenAI-related functionalities were rule creation assistance and virtual assistant support. In most cases, rule creation was a standalone feature, allowing users to either manually define rules, often in languages such as SQL, or generate them with the help of GenAI through natural language prompts. Additionally, 6 of 13 tools integrated GenAI through virtual assistants or copilots. The primary purpose of these copilots was to answer questions about the data, such as “What is the average sales amount for Q4?”, “Which records have missing values in the ‘email’ field?”, or “How many customers are in each state?”, provide descriptive statistics, retrieve relevant data subsets, and generate code or queries. However, in most cases, the generated output, such as a SQL query or a transformation script, had to be manually reviewed and applied by the user in the appropriate interface. Only a few tools, such as Zoho DataPrep and Informatica, supported more direct interaction, where users could issue natural language commands like “Remove all duplicate rows from the dataset” or “Convert all dates into DD/MM/YYYY format, ” and the copilot would execute the task automatically after the output was validated.

Based on this analysis, coupled with discussions with experts, four profiles were identified: GenAI as (1) Translator, (2) Explainer, (3) Resolver, and (4) Integrator. Most observed tools currently fall into the Translator and Explainer categories, meaning GenAI is used to support human decision-making rather than replace it. This observation aligns with feedback from expert interviews, where the potential was recognised, but current implementation and use of GenAI were approached with caution, reflecting concerns about reliability and control.

RQ2. What are the opportunities and limitations of using generative artificial intelligence in data quality management?

Interviewed data quality experts expressed optimism about the potential of GenAI to improve data quality management processes. Core perceived benefits included improved **efficiency**, increased **automation of manual tasks**, and enhanced **scalability** of DQ operations. Experts emphasised that GenAI can significantly reduce the manual effort traditionally required in tasks such as quality rule generation or deduplication.

These results are aligned with existing evidence from previous studies, which have shown that GenAI could reduce manual effort by 30-75%, depending on the use case [53, 54]. For instance, in one scenario, the use of GenAI for data profiling metric suggestions led to an 85% reduction in manual analysis workload [53]. Similarly, employing GenAI to translate natural language inputs into custom SQL queries, with manual validation, was shown to reduce effort by up to 42% [54]. These improvements point not only to time savings but also to **potential reductions in labour costs** over the long term. Although initial implementation costs may be high, the **long-term benefits outweigh early investments** (aligns with [55]).

Ensuring data quality is a complex task that typically demands a combination of domain-specific knowledge, data quality principles, programming skills, and expertise in relevant tools [56]. A recurring theme across interviews was GenAI's ability to help **bridge the gap between business and technical users** - an ability that aligns with the *Translator* and *Explainer* roles in the developed typology. By translating high-level requirements into executable validation logic and explaining technical configurations in accessible terms, GenAI supports more inclusive and accessible data quality workflows [56].

Another strength emphasised by experts was GenAI's **ability to reason contextually and detect patterns or anomalies** that may not be captured by static validation rules. Azeroual [2] similarly notes that models such as GPT-4 are capable of interpreting context by analysing entries in relation to large-scale training data or by cross-referencing internal dataset records. This enables the detection of inconsistencies and anomalies that deviate from expected patterns. This capability aligns with the *Resolver* role, where GenAI is used to detect and resolve data quality issues based on contextual awareness rather than predefined logic alone.

Importantly, the use cases described by the experts aligned closely with the typology conceptualised based on the literature and tool review. Interviewees' thoughts naturally mapped onto the roles of *Translator, Resolver*, *Integrator* and *Explainer* (before the developed typology was presented to them), reinforcing the relevance of these conceptual categories in practice.

However, despite the perceived benefits, several risks and limitations were identified during the interviews. These emerged both from general questions about risks and from more targeted discussions around validating the proposed typology. The most frequently

mentioned concerns that made experts hesitant were the **lack of trust and transparency** in GenAI outputs. These are widely recognised limitations in the literature as well [28, 57, 58, 59, 60].

Experts emphasised that the "black-box" nature poses a challenge in data quality management because users must be able to understand and justify why a particular data transformation or quality rule was applied. This underlines the importance of **explainability**, the model's capability to produce clear explanations or reasoning for its outputs and behaviour, enabling users to better understand and trust the results [60, 61]. If the reasoning behind the AI-generated suggestions is unclear, it undermines accountability and raises concerns related to both **security** and user **complacency**. From a security standpoint, using GenAI for sensitive data introduces risks such as data leakage and the exposure of confidential information to unauthorised parties [2, 61]. Furthermore, such opacity complicates compliance with regulatory frameworks, like the GDPR, which emphasises principles such as transparency, accountability, data minimisation, accuracy, and fairness [63].

Moreover, even the smallest variations in wording, syntax, or the order of examples can lead to significantly different outputs, with users unable to determine why the model responds differently to seemingly similar prompts [64].

While there is extensive research and practical developments on making traditional AI explainable, the literature on generative AI remains limited. This is largely due to the complexity of GenAI models and their reliance on billions of parameters when generating outputs [65]. As a result, tracing the reasoning behind a specific output becomes significantly more difficult [65, 66], making full explainability in GenAI nearly impossible. Nevertheless, certain approaches have been proposed to improve interpretability and provide insights into how these models function. Schneider [65] outlines methods to improve GenAI interpretability, including feature attribution, probing, and self-explanation. Most techniques require white-box access, but some, such as occlusion-based feature attribution or the model's self-explanations, can be applied in black-box settings. However, such post-hoc methods often fall short of revealing the model's actual reasoning processes [67].

Another frequently mentioned risk was **excessive user reliance** on GenAI outputs. Interviewees raised concerns about users accepting AI-generated suggestions without critical evaluation, potentially leading to more errors in the data. These concerns align with what Jason Medd [4] has highlighted as **automation bias** - the tendency to over-rely on automated systems [68] - and **automation complacency**, where users reduce oversight and fail to monitor system outputs adequately [68]. For example, a user may accept a flawed GenAI-generated correction because it appears trustworthy or fail to notice an error due to misplaced confidence in the system's accuracy. In data quality management, such overconfidence can introduce errors instead of resolving them. This connects to the need for GenAI literacy - not just the ability to craft prompts, but a broader competence that includes the ability to critically engage with generative systems, understand their limitations, remain aware of

ethical implications, and reflect on one's own role and potential risks in using them [69, 70]. Both the literature [28, 70, 71] and interviewees stress the importance of maintaining **human oversight** in the loop. This aligns with the principles of **Responsible Generative AI**, which stress transparency, fairness, accountability, and human-centered design, where GenAI augments rather than replaces human judgement [71]. As pointed out in the interviews, GenAI should not be used for the sake of it, in many cases, traditional rule-based systems can remain equally or even more effective. GenAI should be used as a supporter rather than an autonomous decision-maker, providing suggestions but requiring expert validation and accountability.

5.1 Limitations

This study has several limitations. One of the objectives of this thesis was to investigate whether and how current data quality tools on the market integrate GenAI into their functionalities. However, this analysis was limited by restricted access to the tools. Only four tools offered free trials or demo access, and most providers did not respond to email inquiries requesting access for educational purposes. As a result, the evaluation relied primarily on publicly available documentation and website information, which may either overstate GenAI capabilities for marketing purposes or underreport them due to a lack of detail. It is also possible that some relevant tools were inadvertently omitted during the systematic search process. Additionally, some relevant publications may have been missed due to the limitations of the chosen search strings.

Another limitation relates to the sample size of interviewees. While interviews were conducted until thematic saturation was reached, and participants represented a diverse range of professional profiles, it is still possible that certain perspectives or experiences were not captured. The views presented reflect the experiences of the interviewed participants and may not fully represent the broader population of data quality professionals. Future research could benefit from expanding the interview sample size to include a broader range of roles and industries.

Conclusion

This thesis set out to explore the potential opportunities, risks, and practical applications of generative AI in data quality management. Through a combination of tool reviews, expert interviews, and literature analysis, the study explored both the current state of GenAI adoption in data quality tools and the opportunities and limitations of using GenAI in data quality management.

Out of an initial pool of 209 data quality tools, only 13 were found to have already integrated GenAI functionalities, with two more planning to do so in the future. This suggests that while interest in GenAI is increasing, actual implementation remains limited. Based on the tool analysis, a typology of GenAI profiles was developed, identifying four GenAI roles - Translator, Explainer, Resolver and Integrator - and further explored through interviews with data quality professionals. Experts were cautiously optimistic, seeing clear potential for GenAI to improve efficiency, reduce manual workload, and help bridge the gap between technical and business users. At the same time, they raised concerns about the trustworthiness and lack of transparency due to GenAI's black-box nature and the risk of users placing excessive trust in its outputs. Therefore, to integrate GenAI into data quality management in a way that is both practical and responsible, it is essential to maintain human oversight, promote Responsible Generative AI principles, and enhance GenAI literacy.

Overall, the application of GenAI in data quality is still in its early stages. While the potential is promising, several critical risks must be addressed. At this stage, keeping humans in the loop and using GenAI as a supportive partner, rather than a fully autonomous solution, is the key to its safe and effective adoption. Future research could further explore how Responsible GenAI practices can be implemented in data quality management to ensure transparency, accountability, and safe integration of GenAI technologies.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Anastasija Nikiforova, for her ongoing support, guidance, and valuable feedback throughout the thesis. I would also like to thank the interview participants for sharing their time and insights, which were essential to this research.

References

[1] 2025 Outlook: Data Integrity Trends and Insights. Precisely. https://www.precisely.com/resource-center/analystreports/lebow-report-2024 (14.05.2025)

[2] Azeroual O. Can generative AI transform data quality? a critical discussion of ChatGPT's capabilities. *Academia Engineering*, 2024, 1(4). https://doi.org/10.20935/AcadEng7407

[3] Krishnamoorthy N. Data as medicine tech revolutionises healthcare insights: GenAI can help the healthcare sector enhance data quality, enabling real-time analysis and generating synthetic data for improved patient care. *Voice & Data*, 31(8), 2024, 13–15. https://www.voicendata.com/features/data-as-medicine-tech-revolutionises-healthcare-insights-6931181 (08.12.2024)

[4] Medd J. How Will LLMs Impact Data Quality Initiatives? Gartner, 2023. https://emt.gartnerweb.com/ngw/globalassets/en/data-analytics/documents/research-how-will-llms-impact-data-quality-initiatives.pdf

[5] Bobrowski M., Marré M., Yankelevich D. A Homogeneous Framework to Measure Data Quality. *MIT International Conference on Information Quality*, 1999, 115-124. https://www.academia.edu/33150745/A_Homogeneous_Framework_to_Measure_Data_Quality

[6] Wang X., Li X., Xia X. Research on Data Quality Management Methods and Technologies. *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*, 2024, 116–20. https://doi.org/10.1109/ICIPCA61593.2024.10709151

[7] Tayi G. K., Ballou, D. P. Examining data quality. *Communications of the ACM.* New York: Association for Computing Machinery, 1998, 54–57. https://doi.org/10.1145/269012.269021

[8] Wang R. Y., Strong D. M. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 1996, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[9] Zaini N., Seman M. R., Ismail A. N., Majang B. C., Fadhilah S. A. Initiating Data Quality: A Dynamic Rule-Based System for Detecting Errors in Data. *2023 IEEE 11th Conf. Systems, Process & Control (ICSPC)*, 2023, 216-221. https://doi.org/10.1109/ICSPC59664.2023.10420385

[10] How to Improve Your Data Quality. Gartner, 2021. https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality (14.05.2025)

[11] Cichy C., Rass S. An Overview of Data Quality Frameworks. *IEEE Access*, 2019, 7, 2019, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751

[12] Haider K. What is Data Quality Management? A Complete Guide. Astera, 2025. https://www.astera.com/type/blog/data-quality-management/ (13.04.2025)

[13] Allen M., Cervo D. Multi-Domain Master Data Management: Advanced MDM and Data Governance in Practice. Morgan Kaufmann, 2015. https://doi.org/10.1016/C2013-0-18938-6

[14] Nyrhilä P. Improving master data quality in data migration of ERP implementation project. Tampere University of Technology, Master's thesis, 2015. https://urn.fi/URN:NBN:fi:tty-201503231147 (14.05.2025)

[15] Haug A., Zachariassen F., van Liempd D. The costs of poor data quality. *Journal of Industrial Engineering and Management*, 2011, 4(2), 168–93. http://dx.doi.org/10.3926/jiem.2011.v4n2.p168-193

[16] Heinrich B., Hristova D., Klier M., Schiller A., Szubartowicz M. Requirements for Data Quality Metrics. *Journal of Data and Information Quality (JDIQ)*, 2018, 9, 1-32. https://doi.org/10.1145/3148238

[17] Batini C., Scannapieco M. Data Quality: Concepts, Methodologies and Techniques. Berlin: Springer, 2006. ttps://doi.org/10.1007/3-540-33173-5

[18] Wand Y., Wang, R. Y. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 1996, 39(11), 86–95. https://doi.org/10.1145/240455.240479

[19] Ballou D. P., Pazer H. L. Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 1985, 31(2), 150–162. https://doi.org/10.1287/mnsc.31.2.150

[20] Kilani S. O., Amao I. K., Ojo N.A., Samson P. A. Ai and Machine Learning-Powered Automated Data Cleaning Methods: Improving Data Quality. *International Journal of Advance Research Publication and Reviews,* 2025, 2(4), 35-47. https://ijarpr.com/uploads/V2ISSUE4/IJARPR0603.pdf

[21] Thirunagalingam A. AI-Powered Continuous Data Quality Improvement: Techniques, Benefits, and Case Studies. *3rd International Conference on Research in Multidisciplinary Studies*, 2024, 10, 38-46. https://dx.doi.org/10.2139/ssrn.5047709

[22] Banerjee A. Automating Data Engineering Workflows with AI and Machine Learning. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2024, 5(2), 9–16. https://doi.org/10.63282/3050-9262.IJAIDSML-V5I2P102

[23] Sengar S. S., Hasan A. B., Kumar S., Carroll F. Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications*, 2024. https://doi.org/10.1007/s11042-024-20016-1

[24] Takale D, Mahalle P, Sule B. Advancements and Applications of Generative Artificial Intelligence. *Journal of Information Technology and Sciences*, 2024, 10, 20–7. https://matjournals.net/engineering/index.php/JOITS/article/view/188

[25] Ooi K. B., Tan G. W. H., Al-Emran M., Al-Sharafi M. A., Capatina A., Chakraborty A., Dwivedi Y. K., Huang T.-L., Kar A. K., Lee V. H., Loh X. M., Micu A., Mikalef P., Mogaji E., Pandey N., Raman R., Rana N. P., Sarker P., Sharma A., Teng C.-I., Fosso Wamba S., Wong L.-W. The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. *Journal of Computer Information Systems*, 2025, 65(1), 76–107. https://doi.org/10.1080/08874417.2023.2261010

[26] Cooper G. Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *Journal of Science Education and Technology*, 2023, 32, 444-452. https://doi.org/10.1007/s10956-023-10039-y

[27] IBM. What are large language models (LLMs)? IBM, 2023. https://www.ibm.com/think/topics/large-language-models (13.04.2025)

[28] Fui-Hoon Nah F., Zheng R., Cai J., Siau K., Chen L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*. 2023, 25(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

[29] Banh L., Strobel G. Generative artificial intelligence. *Electron Markets*. 2023, 33(63). https://doi.org/10.1007/s12525-023-00680-1

[30] Ji Z., Lee N., Frieske R., Yu T., Su D., Xu Y., Ishii E., Bang Y., Chen D., Dai W., Chan H. S., Madotto A., Fung P. Survey of Hallucination in Natural Language Generation. Association for Computing Machinery. *ACM Comput Surveys*, 2023, 55(12), 1–38. http://dx.doi.org/10.1145/3571730

[31] Law M. Generative AI and ML fuelling a revolution in data quality. AI Magazine, 2023. https://aimagazine.com/articles/generative-ai-and-ml-fuelling-a-revolution-in-data-quality (6.05.2025)

[32] Jamaluddin J., Gaffar N. A., Din N. S. S. Hallucination: A key challenge to Artificial Intelligence-Generated writing. *Malays Fam Physician*, 2023, 18, 68. https://doi.org/10.51866/lte.527

[33] Varma S., Shivam S., Ray B., Biswas S. Reimagining enterprise data management using generative artificial intelligence. *Proceedings of the 11th IEEE Swiss Conference on Data Science (SDS),* 2024, 107-114. https://doi.org/10.1109/SDS60720.2024.00023

[34] Liu I., Wongsosaputro M. Compliance redefined: Using GenAI to navigate a complex regulatory landscape with reduced risks and costs. *Journal of Digital Banking*, 2024, 313–322. https://doi.org/10.69554/LGTJ4498

[35] Ehrlinger L, Wöß W. A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data*, 2022, 5. https://doi.org/10.3389/fdata.2022.850611

[36] Martinsaari H. C. Toward an Automated Data Quality Rule Detection in Data Warehouses. University of Tartu, Institute of Computer Science, Master's Thesis. 2023. https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=77161 (14.05.2025)

[37] Zhou Y., Tu F., Sha K., Ding J., Chen H. A survey on data quality dimensions and tools for machine learning. *IEEE International Conference on Artificial Intelligence Testing (AITest)*, 2024, 120-131. https://doi.org/10.1109/AITest62860.2024.00023

[38] Kitchenham B., Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report. 2007. https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf

[39] Brereton P., Kitchenham B. A., Budgen D., Turner M., Khalil M. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 2007, 80(4), 571–83. https://doi.org/10.1016/j.jss.2006.07.009

[40] Andmekaitse ja infoturbe leksikon. Cybernetica AS. https://akit.cyber.ee/

[41] Jaadi Z. Data Standardization: How to Do It and Why It Matters. Built In, 2025. https://builtin.com/data-science/when-and-why-standardize-your-data (13.04.2025)

[42] IBM. What is data integration? https://www.ibm.com/think/topics/data-integration (13.04.2025)

[43] OpenAI (2025). ChatGPT (version GPT-4). https://chat.openai.com

[44] Barnard J., Stryker C. What is anomaly detection? IBM Think, 2023. https://www.ibm.com/think/topics/anomaly-detection (13.04.2025)

[45] What is data quality monitoring? IBM, 2023. https://www.ibm.com/think/topics/data-quality-monitoring-techniques (13.04.2025)

[46] Data management glossary. SAP, 2022. https://www.sap.com/resources/data-management-glossary (13.04.2025)

[47] Stapley E., O'Keeffe S., Midgley N. Developing Typologies in Qualitative Research: The Use of Ideal-type Analysis. *International Journal of Qualitative Methods*, 2022, 21. https://doi.org/10.1177/16094069221100633

[48] Lehmann S. L., Dahlke J., Pianta V., Ebersberger B. Artificial intelligence and corporate ideation systems. *Journal of Product Innovation Management,* 1–26. https://doi.org/10.1111/jpim.12782

[49] Guest G., Bunce, A., Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, 2006, 18(1), 59–82. https://doi.org/10.1177/1525822X05279903

[50] Malterud K., Siersma V. D., Guassora A. D. Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, 2016, 26(13), 1753–1760. https://doi.org/10.1177/1049732315617444

[51] Mason M. Sample Size and Saturation in PhD Studies Using Qualitative Interviews. *Forum Qualitative Sozialforschung Forum: Qualitative Social Research*, 2010, 11(3). https://doi.org/10.17169/fqs-11.3.1428

[52] Braun V., Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 2006, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[53] Varma S., Shivam S., Ray B., Biswas S. Reimagining Enterprise Data Management using Generative Artificial Intelligence. *11th IEEE Swiss Conference on Data Science (SDS)*, 2024, 107–14. https://doi.org/10.1109/SDS60720.2024.00023

[54] Singh A. Shetty A., Ehtesham A., Kumar S., Khoei T. T. A Survey of Large Language Model-Based Generative AI for Text-to-SQL: Benchmarks, Applications, Use Cases, and Challenges. 2025. https://doi.org/10.48550/arXiv.2412.05208

[55] Chen Y. Analytical Study on Revolutionizing Data Transformation with Generative AI in Data Engineering. *International Journal of Unique and New Updates*, 2019, 1(1), 34-41. https://ijunu.com/index.php/journal/article/view/5

[56] Inala J. P., Wang C., Drucker S., Ramos G., Dibia V., Riche N., Brown D., Marshall D., Gao J. Data Analysis in the Era of Generative AI. arXiv, 2024. https://doi.org/10.48550/arXiv.2409.18475

[57] Manduchi L., Pandey K., Meister C., Bamler R., Cotterell R., Däubener S., Fellenz S., Fischer A., Gärtner T., Kirchler M., Kloft M., Li Y., Lippert C., de Melo G., Nalisnick E., Ommer B., Ranganath R., Rudolph M., Ullrich K., Van den Broeck G., Vogt J. E., Wang Y., Wenzel F., Wood F., Mandt S., Fortuin V. On the Challenges and Opportunities in Generative AI. ArXiv, 2024. https://doi.org/10.48550/arXiv.2403.00025

[58] Routray S. K., Jha M. K., Sharmila K. P., Javali A., Pappa M., Singh M. Generative Artificial Intelligence: Principles, Potentials and Challenges. *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, 2024, 211-216. https://doi.org/10.1109/ICSSAS64001.2024.10760992

[59] Al-kfairy M., Mustafa D., Kshetri N., Insiew M., Alfandi O. Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective. *Informatics,* 2024, 11(3), 58. https://doi.org/10.3390/informatics11030058

[60] Vyas P., Vyas G. Generative Artificial Intelligence: Current Trends, Issues, and Challenges. *IT Professional*. 2025, 27, 20-26. https://doi.ieeecomputersociety.org/10.1109/MITP.2024.3516058

[61] Liao Q. V., Varshney K. R. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv, 2022. https://doi.org/10.48550/arXiv.2110.10790

[62] Goriparthi S. Tracing data lineage with generative AI: Improving data transparency and compliance. *International Journal of Artificial Intelligence & Machine Learning*, 2023, 155-165. https://lib-index.com/index.php/IJAIML/article/view/IJAIML_02_01_015

[63] Athanasopoulou D.-D. Data protection in the era of generative artificial intelligence: navigating GDPR compliance challenges in medical applications of ChatGPT. National and Kapodistrian University of Athens, Law school, postgraduate thesis. 2024. https://pergamos.lib.uoa.gr/uoa/dl/object/3449277

[64] Kaddour J., Harris J., Mozes M., Bradley H., Raileanu R., McHardy R. Challenges and Applications of Large Language Models. arXiv, 2023. https://doi.org/10.48550/arXiv.2307.10169

[65] Schneider J. Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda. *Artif Intell Rev*, 2024, 57. https://doi.org/10.1007/s10462-024-10916-x

[66] Davies A. Explainable AI – Making Your Gen AI Understandable. LinkedIn, 2025.

https://www.linkedin.com/pulse/explainable-ai-making-your-gen-understandable-adam-davies-pzf7e/ (8.05.2025)

[67] Dittmar L. The Explainability Challenge of Generative AI and LLMs. OCEG. https://www.oceg.org/the-explainability-challenge-of-generative-ai-and-llms/ (8.05.2025)

[68] Goddard K., Roudsari A., Wyatt J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 2012, 19(1), 121-127. https://doi.org/10.1136/amiajnl-2011-000089

[69] Zhang C., Magerko B. Generative AI Literacy: A Comprehensive Framework for Literacy and Responsible Use. arXiv, 2025. https://doi.org/10.48550/arXiv.2504.19038

[70] Cox A. Algorithmic Literacy, AI Literacy and Responsible Generative AI Literacy. *Journal of Web Librarianship*, 2024, 18(3), 93–110. https://doi.org/10.1080/19322909.2024.2395341

[71] Raza S., Qureshi R., Zahid A., Fioresi J., Sadak F., Saeed M., Sapkota R., Jain A., Zafar A., Ul Hassan M., Zafar A., Maqbool H., Vayani A., Wu J., Shoman M. Who is Responsible? The Data, Models, Users or Regulations? A Comprehensive Survey on Responsible Generative AI for a Sustainable Future. arXiv, 2025. https://doi.org/10.48550/arXiv.2502.08650

Appendix A - Interview Questions

Thank you for participating in this study on the use of generative artificial intelligence in data quality management. This study is conducted by Karen Roht under the supervision of Dr. Anastasija Nikiforova at the University of Tartu. The aim of this study is to explore the potential applications of generative artificial intelligence in data quality management. To achieve this, a systematic analysis of existing data quality tools has been conducted to identify where generative AI is already being utilised. Your insights will help validate the findings of this analysis and offer a deeper understanding of the opportunities and challenges associated with integrating generative AI into data quality management.

General Profile

- Job position/title
- Academic qualification/title
- Expertise / professional background (please indicate how your background is related to data quality management)
- Have you worked with or implemented generative AI solutions for data quality management before?
- Work experience (in years)

Current Situation

- What do you consider to be the biggest challenges in data quality management today?
- Which aspects of data quality are currently the most resource-intensive or difficult to maintain?

Potential Uses of Generative AI

- In your opinion, which specific tasks in data quality management could be improved with generative AI? How do you think generative AI could help address these challenges?
- What do you see as the main benefits of using generative AI for data quality management?

In addition to your own experiences and knowledge, I would like you to review the typology I developed for leveraging generative AI in data quality management. I conducted a systematic analysis of data quality tools and identified where they incorporate generative AI. Based on this, I have created four ideal types. I would greatly appreciate it if you could validate or invalidate these types based on your experience. Do you find these categories to be accurate? I would like you to:

1) **Suggest practical examples that would enhance the types.**
2) **Identify the benefits associated with each type.**
3) **Identify the challenges associated with each type.**

1. **GenAI as a translator**
   Pains: Ever-evolving business requirements and the need for technical expertise in defining and implementing data quality rules.
   Pain reliever: GenAI allows users to define data quality checks in natural language and converts them into quality rules.

	Practical examples	Benefits	Risks
GenAI as a translator			

2. **GenAI as an explainer**
   Pains: Users often struggle to understand why specific issues occur, due to gaps in expertise or a lack of transparency in data processing. Time-consuming root-cause analysis.
   Pain reliever: GenAI translates complex data quality analysis results into simple, natural language. Users can ask about specific data quality issues. Identifies potential root causes and explains why these issues are arising.

	Practical examples	Benefits	Risks
GenAI as an explainer			

3. **GenAI as a resolver**
   Pains: Identifying and fixing data quality issues manually can be slow and error-prone. Manually analysing datasets becomes increasingly difficult as datasets grow.
   Pain reliever: GenAI automates the issue identification and fixing process. Generates suggested fixes based on patterns. For example, the user can ask GenAI to fix inconsistent currency formatting.

	Practical examples	Benefits	Risks
GenAI as a resolver			

4. **GenAI as a data integrator**
   Pains: Data integration across multiple systems is complex due to schema mismatches and inconsistent data structures.
   Pain reliever: GenAI assists in aligning and integrating data across different systems and schemas. For instance, users can request GenAI to generate data integration mappings.

	Practical examples	Benefits	Risks
GenAI as a data integrator			

Risks and Limitations

- What potential risks or downsides do you see in GenAI-empowered data quality management?
- Are there any data quality management tasks where you believe generative AI should not be used? Why?

Appendix B - Relevant Academic Publications From the Literature Review

This table provides an overview of all relevant academic publications selected during the literature search. The tools mentioned here were carried forward into the initial screening phase.

SQ1: Web of Science, “DATA QUALITY” AND “TOOL”
SQ2: Scopus, "DATA QUALITY TOOL"
SQ3: Scopus, TITLE-ABS-KEY (“DATA QUALITY” AND TOOL)

	Title	**Authors**	**Year**	**Source Title**	**Mentioned tools**
SQ1, SQ2	Extending Achilles Heel Data Quality Tool with New Rules Informed by Multi-Site Data Quality Comparison	Huser V., Li X., Zhang Z., Jung S., Park R. W., Banda J., Razzaghi H., Londhe A., Natarajan, K.	2019	Studies in health technology and informatics	Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES)

	Title	Authors	Year	Source Title	Mentioned tools
SQ1, SQ2, SQ3	A Survey of Data Quality Measurement and Monitoring Tools	Ehrlinger L., Wöß W.	2022	Front Big Data	Aggregate Profiler, Apache Griffin, Ataccama ONE, DataCleaner, Datamartist, Experian Pandora, Informatica Data Quality, IBM InfoSphere Information Server for Data Quality, InfoZoom, MobyDQ, OpenRefine, Oracle Enterprise Data Quality, Talend Open Studio for Data Quality, SAS Data Quality
SQ1, SQ3	DaQL 2.0: Measure Data Quality based on Entity Models	Lettner C., Stumptner R., Fragner W., Rauchenzauner F., Ehrlinger L.	2021	Procedia Computer Science	DaQL
SQ1	DQAgui: a graphical user interface for the MIRACUM data quality assessment tool	Mang J., Seuchter S., Gulden C., Schild S., Kraska D., Prokosch H., Kapsner L. A.	2022	BMC MEDICAL INFORMATICS AND DECISION MAKING	DQAgui
SQ1, SQ2, SQ3	A Survey on Data Quality Dimensions and Tools for Machine Learning	Zhou Y., Tu F., Sha K., Ding J., Chen H.	2024	2024 IEEE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE TESTING	Kylo, MobyDQ, Apache Griffin, SQL Power Architect, Aggregate Profiler, YData Quality, DataCleaner, WinPure, SQL PowerDQguru, Deequ, Dataedo, OpenRefine, Great Expectations, Soda Core, Ataccama

	Title	Authors	Year	Source Title	Mentioned tools
					ONE, whylogs, Evidently
SQ1, SQ3	Data Profiling for Data Quality Improvement with Openrefine	Kusumasari T., Fitria	2016	PROCEEDINGS OF 2016 INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY SYSTEMS AND INNOVATION	OpenRefine, IBM Infosphere Information Analyzer, Oracle Enterprise Data Quality, Talend Data Quality, SAP Business Objects Data Insight & SAP Business Objects Information Steward, Informatica Data Explore
SQ1, SQ3	DQ-MAN: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems	Buelvas J., Múnera D., Gaviria N.	2023	INTERNET OF THINGS	DQ-MAN
SQ1, SQ3	TAQIH, a tool for tabular data quality assessment and improvement in the context of health data	Sánchez R., Iraola A., Unanue G., Carlin P.	2019	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	TAQIH
SQ1, SQ2, SQ3	Automating Data Quality Monitoring with Reference Data Profiles	Ehrlinger L., Werth B., Wöss W.	2023	DATA MANAGEMENT TECHNOLOGIES AND APPLICATIONS	DQ-MeeRKat, Oracle Enterprise Data Quality, SAS, Talend, Informatica, HoloDetect
SQ1	GazeMetrics: An Open-Source Tool for Measuring the Data Quality of HMD-based Eye Trackers	Adhanom I., Lee S., Folmer E., MacNeilage P.	2020	ETRA 2020 SHORT PAPERS: ACM SYMPOSIUM ON EYE TRACKING RESEARCH & APPLICATIONS	GazeMetrics

	Title	Authors	Year	Source Title	Mentioned tools
SQ1, SQ2	DQ-MeeRKat: Automating Data Quality Monitoring with a Reference-Data-Profile-Annotated Knowledge Graph	Ehrlinger L., Gindlhumer A., Huber L., Wöss W.	2021	PROCEEDINGS OF THE 10TH INTERNATIONAL CONFERENCE ON DATA SCIENCE, TECHNOLOGY AND APPLICATIONS (DATA)	DQ-MeeRKat
SQ1, SQ3	A Data-centric AI Framework for Automating Exploratory Data Analysis and Data Quality Tasks	Patel H., Guttula S., Gupta N., Hans S., Mittal R., Lokesh N.	2023	ACM JOURNAL OF DATA AND INFORMATION QUALITY	Deequ, Pandas Profiling, TensorFlow Validation, TDDA, Great Expectations
SQ1	EasyQC: Tool with Interactive User Interface for Efficient Next-Generation Sequencing Data Quality Control	Rangamaran V., Uppili B., Gopal D., Ramalingam K.	2018	JOURNAL OF COMPUTATIONAL BIOLOGY	EasyQC
SQ1, SQ3	QualiBD: A Tool for Modelling Quality Requirements for Big Data Applications	Arruda D., Madhavji N.	2019	2019 IEEE INTERNATIONAL CONFERENCE ON BIG DATA	QualiBD
SQ1	Improved Data Accuracy Assessment Tool for Information Management Systems	Maziku H.	2020	2020 6TH INTERNATIONAL CONFERENCE ON COMMUNICATION AND INFORMATION PROCESSING	Data Accuracy Assessment Tool, Routine Data Quality Assessment Tool, Data Verification and Improvement Tool

	Title	Authors	Year	Source Title	Mentioned tools
SQ1, SQ3	CLEAN4TSDB: A Data Cleaning Tool for Time Series Databases	Ding X., Song Y., Wang H., Yang D., Wang C., Wang J.	2024	PROCEEDINGS OF THE VLDB ENDOWMENT	Clean4TSDB
SQ1, SQ3	Interactive Data Mashups for User-Centric Data Analysis	Behringer M., Hirmer P.	2023	35TH INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT	PowerBI, Knime, RapidMiner
SQ1, SQ3	OmicsEV: a tool for comprehensive quality evaluation of omics data tables	Wen B., Jaehnig E., Zhang B.	2022	BIOINFORMATICS	OmicsEV
SQ1, SQ3	DataPilot: Utilizing Quality and Usage Information for Subset Selection during Visual Data Preparation	Narechania A., Du F., Sinha A. R., Rossi R., Hoffswell J., Guo S.	2023	PROCEEDINGS OF THE 2023 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS	DataPilot, OpenRefine
SQ1, SQ3	Data Preparation: A Survey of Commercial Tools	Hameed M., Naumann F.	2020	SIGMOD RECORD	Altair Monarch Data Preparation, SAP Agile Data Preparation, SAS Data Preparation, Tableau Prep, Talend Data Preparation, Trifacta Wrangler
SQ1, SQ3	Data-Debugging Through Interactive Visual Explanations	Afzal S., Chaudhary A., Gupta N., Patel H., Spina C., Wang	2021	TRENDS AND APPLICATIONS IN KNOWLEDGE	Wrangler, OpenRefine, explAIner, TELEGAM, Tensorflow Embedding Projector

	Title	Authors	Year	Source Title	Mentioned tools
		D.		DISCOVERY AND DATA MINING	
SQ1, SQ3	A Comparative Study of Data Cleaning Tools	Oni S., Chen Z., Hoban S., Jademi O.	2019	INTERNATIONAL JOURNAL OF DATA WAREHOUSING AND MINING	OpenRefine, DataWrangler
SQ1	Automatic Assessment of Quality of your Data for AI	Patel H., Gupta N., Panwar N., Mittal R., Mehta S., Guttula S., Mujumdar S., Afzal S., Bedathur S., Munigala V.	2022	PROCEEDINGS OF THE 5TH JOINT INTERNATIONAL CONFERENCE ON DATA SCIENCE & MANAGEMENT OF DATA	Pandas Profiling, Deepqu, IBM's Data Quality for AI,
SQ1	CleanCloud: Cleaning Big Data on Cloud	Wang H., Ding X., Chen X., Li J., Gao H.	2017	CIKM'17: PROCEEDINGS OF THE 2017 ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT	CleanCloud, BigDansing
SQ1, SQ3	Detangler: Helping Data Scientists Explore, Understand, and Debug Data Wrangling Pipelines	Shrestha N., Chopra B., Henley A., Parnin C.	2023	2023 IEEE SYMPOSIUM ON VISUAL LANGUAGES AND HUMAN-CENTRIC	Detangler

	Title	Authors	Year	Source Title	Mentioned tools
				COMPUTING	
SQ1, SQ3	FASTOD: Bringing Order to Data	Mihaylov A., Godfrey P., Golab L., Kargar M., Srivastava D., Szlichta J.	2018	2018 IEEE 34TH INTERNATIONAL CONFERENCE ON DATA ENGINEERING	FASTOD
SQ1	Relating instance hardness to classification performance in a dataset: a visual approach	Paiva P., Moreno C., Smith-Miles K., Valeriano M., Lorena A.	2022	MACHINE LEARNING	PyHard, MATILDA (Melbourne Algorithm Test Instance Library with Data Analytics)
SQ1	From Queriability to Informativity, Assessing "Quality in Use" of DBpedia and YAGO	Ruan T., Li Y., Wang H., Zhao L.	2016	SEMANTIC WEB: LATEST ADVANCES AND NEW DOMAINS	Kbmetrics
SQ1. SQ2. SQ3	How to Inspect and Measure Data Quality about Scientific Publications: Use Case of Wikipedia and CRIS Databases	Azeroual O., Lewoniewski W.	2020	ALGORITHMS	DataCleaner

	Title	Authors	Year	Source Title	Mentioned tools
SQ1, SQ2	Stream2segment: An Open-Source Tool for Downloading, Processing, and Visualizing Massive Event-Based Seismic Waveform Datasets	Zaccarelli R., Bindi D., Strollo A., Quinteros J., Cotton F.	2019	SEISMOLOGICAL RESEARCH LETTERS	Stream2segment
SQ2, SQ3	Data Quality Assessment in the Wild: Findings from GitHub	Ustunboyacioglu I., Kumara I., Di Nucci D., Tamburri D. A., Van Den Heuvel W. J.	2024	ACM International Conference Proceeding Series	Deequ, PyDeequ, Pandera, Great Expectations, Tensorflow Data Validation
SQ2, SQ3	A Functional Taxonomy of Data Quality Tools: Insights from Science and Practice	Altendeitering M., Tomczyk M.	2022	17th International Conference on Wirtschaftsinformatik	Ataccama, Experian, IBM, Infogix, Informatica, InfoZoom, Innovative Systems, Melissa, MIOSoft, nModal Solutions, Oracle, Precisely, Redpoint, SAP, SAS, Syniti, Talend
SQ2, SQ3	Data Cleansing Processing using Pentaho Data Integration: Case Study Data Deduplication	Setyawan D. C., Kusumasari T. F., Alam E. N.	2020	Proceedings - 2020 6th International Conference on Science and Technology	Pentaho Data Integration
SQ2, SQ3	Open source data quality tools: Revisited	Wen B., Jaehnig E., Zhang B.	2016	Advances in Intelligent Systems and Computing	Talend Open Studio, DataCleaner WinPure, Data Preparator, Data Match, DataMartist, Pentaho Data

	Title	Authors	Year	Source Title	Mentioned tools
					Integration (Formerly Pentaho Kettle), SQL Power Architect, SQL Power DQguru, DQAnalyzer
SQ2	A model-based evaluation of data quality activities in KDD	Mezzanzanica M., Boselli R., Cesarini M., Mercorio F.	2015	Information Processing and Management	NADEEF
SQ3	KGHeartBeat: An Open Source Tool for Periodically Evaluating the Quality of Knowledge Graphs	Pellegrino M. A., Rula A., Tuozzo G.	2025	Lecture Notes in Computer Science	KGHeartBeat, Sieve, RDFUnit, LiQuate, DaCura, LinkQA, SPARQLES, Loupe API, LD Sniffer, SemQuire, DYLDO DistQualityAssessment, LODLaundromat, Luzzu, ABECTO, Roomba, YummyData
SQ3	LLMClean: Context-Aware Tabular Data Cleaning via LLM-Generated OFDs	Biester F., Abdelaal M., Del Gaudio D.	2025	Communications in Computer and Information Science	LLMClean, Baran, HoloClean, ED2, Pandas' Missing Value Detector (MVD), Raha, dBoost, Nadeef,
SQ3	Giraffe: A tool for comprehensive processing and visualization of multiple long-read sequencing data	Liu X., Shao Y., Guo Z., Ni Y., Sun X., Leung A. Y. H.	2024	Computational and Structural Biotechnology Journal	Giraffe
SQ3	Data quality assurance in research data repositories: a theory-guided exploration and model	Stvilia B., Lee D. J.	2024	Journal of Documentation	OpenRefine

	Title	Authors	Year	Source Title	Mentioned tools
SQ3	Enhancing Data Trustworthiness in Explorative Analysis: An Interactive Approach for Data Quality Monitoring	Behringer M., Hirmer P., Villanueva A., Rapp J., Mitschang B.	2024	SN Computer Science	Knime, RapidMiner, Tableau
SQ3	LinkedDataOps:quality oriented end-to-end geospatial linked data production governance	Yaman B., Thompson K., Fahey F., Brennan R.	2024	Semantic Web	1Spatial 1Integrate, Luzzu
SQ2, SQ3	Cleenex: Support for User Involvement during an Iterative Data Cleaning Process	Pereira J. L. M, Fonseca M. J., Lopes A., Galhardas H.	2024	Journal of Data and Information Quality	OpenRefine, Precisely Data Integrity Suite, Pentaho Data Integration, NADEEF, Holoclean, ActiveClean, CoClean, Baran, ICARUS, FALCON, Cleenex
SQ3	SAGED: Few-Shot Meta Learning for Tabular Data Error Detection	Abdelaal M., Ktitarev T., Städtler D., Schöning H.	2024	Advances in Database Technology - EDBT	RAHA, OpenRefine, Trifacta, AutoCure, Nadeef, HoloClean, HoloDetect, ED2, Saged, KATARA, dBoost, min-K, FAHES
SQ3	MCLIENT – A WEB TOOLKIT FOR OPEN DATA PUBLISHING	Wenige L., Stadler C., Frank C. W., Martin M., Figura R.	2024	gis.Science - Die Zeitschrift fur Geoinformatik	MClient, CKAN, DKAN
SQ3	Data Quality Optimization for Decision Making Using Ataccama Toolkit: A Sustainable Perspective	Jamal A., Quadri M. P., Rafeeq M.	2023	International Journal on Recent and Innovation Trends in Computing and Communication	Ataccama ONE, Informatica Multidomain MDM, SAP Master Data Governance, IBM InfoSphere Master Data Management, Innovative Systems, Semarchy xDM,

	Title	Authors	Year	Source Title	Mentioned tools
					Integrate.io, Altova MapForce, Talend, Pentaho, OpenText
SQ3	Lumigi: Shining Light on Your Process Data	Vugs L., van Asseldonk M., van Son N.	2021	CEUR Workshop Proceedings	Lumigi
SQ2, SQ3	Data cleaning techniques in detecting tendencies in software engineering	Georgieva P., Nikolova E., Orozova D.	2020	2020 43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020 - Proceedings	Informatica PowerCenter, IBM Information Server, Talend Open Studio, Oracle Data Integrator, SQL Server Integration Services, SAS Data Integration Studio, Pentaho Data Integration, Clover ETL, CleenexQCs, LlunaticEgds, Nadeef, Eracer
SQ3	A Study on the Aspects of Quality of Big Data on Online Business and Recent Tools and Trends towards Cleaning Dirty Data	Hossen M. I., Goh M., Hossen A., Rahman M. A.	2020	2020 11th IEEE Control and System Graduate Research Colloquium, ICSGRC 2020 - Proceedings	SP Data Services, Infosphere Quality Stage, Data Quality management (SAS), Global data quality suite
SQ3	Data cleaning: A case study with openrefine and trifacta wrangler	Petrova-Antonova D., Tancheva R.	2020	Communications in Computer and Information Science	WinPure Clean & Match, RapidMiner, TIBCO Clarity, Data Ladder, OpenRefine, Trifacta, Data Cleaner
SQ3	Time Series Data Cleaning: A Survey	Wang X., Wang C.	2020	IEEE Access	PIClean, HoloClean, ActiveClean, Cleanits, MLClean, ASAP, EDCleaner, PACAS, TSOutlier

	Title	Authors	Year	Source Title	Mentioned tools
SQ3	Data quality in ETL process: A preliminary study	Souibgui M., Atigui F., Zammali S., Cherfi S., Yahia S. B.	2019	Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019	MassEETL, POIESIS, Talend Data Integration, Pentaho Data Integration Microsoft SQL Server Integration Services, Informatica Data Integration
SQ1, SQ3	Big data validation case study	Xie C., Gao J., Tao C.	2017	Proceedings - 3rd IEEE International Conference on Big Data Computing Service and Applications, BigDataService 2017	DataCleaner, Datameer, Tableau, Pentahao Advanced ETL Processor, Talend, Zoho, Query surge, Splunk
SQ3	Big data validation and quality assurance - IssuSes, challenges, and needs	Gao J., Xie C., Tao C.	2016	Proceedings - 2016 IEEE Symposium on Service-Oriented System Engineering	Datameer, Talend Open Studio, Informatica, IBM, QuerySurge, C3 Integrity, Microsoft Azure HDInsight, SAP HANA, Jumbune

Appendix C - Relevant Articles From the Literature Review

This table provides an overview of all relevant articles selected from Google during the literature search. The tools mentioned here were carried forward into the initial screening phase.

Title	URL	Perspective	Mentioned tools
16 top data governance tools to know about in 2025	link	Technology media and marketing	Alation Data Governance, Apache Atlas, Ataccama One, Collibra Data Governance, Erwin Data Intelligence by Quest, IBM Cloud Pak for Data, Informatica Cloud Data Governance and Catalog, OneTrust Data Discovery & Classification, Oracle Enterprise Data Management, Precisely Data Integrity Suite, Rocket Data Intelligence, Rocket Data Intelligence, SAP Master Data Governance, SAS Information Governance, Semarchy Data Intelligence, Syniti Knowledge Platform, Talend Data Fabric
Best customer data platforms in 2025	link	Technology news site	Hubspot
The Ultimate Guide to Data Quality Tools: Top Solutions, Features, and Selection Criteria for 2025	link	Online education platform	Informatica Data Quality, Talend Data Quality, IBM InfoSphere, SAS Data Quality, Great Expectations, Apache Griffin
Top Data Quality (DQ) Tools	link	Technology research and review platform	PiLog Data Quality Management, SAS Data Quality, Talend Data Quality, Alteryx, Zoominfo OperationsOS, Microsoft Data Quality Services, Informatica Data Quality, Melissa Data Quality Suite, Oracle Enterprise Data Quality, Aperture Data Studio, SAP Data Services, Validity DemandTools, Ataccama ONE Platform, IBM InfoSphere QualityStage, Cloudingo
Gartner Magic	link	Research	Melissa, Datactics, MIOsoft, Redpoint,

Title	URL	Perspective	Mentioned tools
Quadrant for Data Quality: Evaluation & Ranking Criteria		and advisory firm	Innovative Systems, Experian, Collibra, Syniti TIBCO Software, Precisely, Ataccama, SAS Talend, SAP, IBM, Informatica
The 9 Best Augmented Data Quality Tools and Software for 2025	link	Technology news site	Ataccama ONE, Collibra, Informatica, Innovative Systems, Oracle, Precisely, SAP, Syniti, Talend
12 Best Data Quality Tools for 2025	link	Open-source data version control system platform	Great Expectations, Deequ, Monte Carlo, Anomalo, Lightup, Bigeye, Acceldata, Observe.ai, Datafold, Collibra, dbt Core, Soda Core
The Top 10 Data Quality Tools	link	Technology research and review platform	Ataccama Data Quality & Governance, Collibra Data Quality & Observability, Experian Aperture Data Studio, IBM InfoSphere Information Server for Data Quality, Informatica Cloud Data Quality, Melissa Unison, Precisely Data Integrity Suite, SAP Master Data Governance, SAS Viya, Talend Data Quality
Top Data Quality Tools You Should Know in 2025	link	Online education platform	D&B Connect, Syncari, Duplicate Check for Salesforce, SAS Data Quality, DQE One,
Top Data Quality Tools for 2025: Improve Data Quality	link	Online education platform	OpenRefine, Talend, Cloudingo, IBM InfoSphere, Data Ladder, Ataccama ONE, Experian, Oracle, Syniti, SAS
Top 10 Best Data Quality Tools for 2024	link	Technology news site	Ab Initio, SAS Data Quality, DQLabs Platform, OpenRefine, Precisely Data Integrity Suite, Oracle Enterprise Data Quality, Talend Data Fabric, SAP Data Services, Ataccama ONE, Informatica Cloud Data Quality
A guide to open-source data quality tools in late 2023	link	Publishing platform	Unravel, Collibra, acceldata, timeseer.ai, Soda, Anomalo, Monte Carlo, Datafold, metaplane, Sync, Bigeye, Precisely, Validio, Manta, Databand, Sifflet, Talend, Great Expectations, lightup

Title	URL	Perspective	Mentioned tools
Top 7 Data Quality Tools and Software	link	Publication and community platform	Data Ladder, OpenRefine, Talend, Ataccama, Dataedo, Precisely, Informatica
Top 5 Data Quality Tools in 2023	link	Technology news site	Cleanlab Studio, Informatica, SAS Data Quality, Deequ, OpenRefine
5 Best Data Quality Tools for Your Use in 2023	link	Online education platform	Ataccama, Informatica, Innovative Systems, Oracle, Precisely
State of Data Quality Tools 2024 Q1	link	AI and data consultancy company	re_data, Elementary, Soda, Great Expectations
5 Data Quality Tools to Ensure Accuracy and Integrity	link	AI and data analytics community portal	Data Ladder, OpenRefine, Tibco Clarity, Trifacta, Winpure and tray.io, experian, Infosys, iCEDQ, Claravine

Appendix D - Licence

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Karen Roht ,
(author´s name)

grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

Generative AI in Data Quality Management ,
(title of thesis)

supervised by Anastasija Nikiforova ;
(supervisor´s name)

grant the University of Tartu the permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from *14/05/2025* until the expiry of the term of copyright;
am aware that the author retains the rights specified in points 1 and 2;
confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Karen Roht
14/05/2025