SUPPLIER MASTER DATA INTEGRATION IN PHILIPS

MASTER THESIS:

BUSINESS INFORMATION TECHNOLOGY

DYLAN YE | S1495100

Supervisors:

Dr. F.A. Bukhsh

Dr. A. Abhishta

N. Mitchell (Philips)

SEPTEMBER 2023

UNIVERSITY OF TWENTE.

Table of Contents

List of Figures

List of Tables

Executive summary

[**Context/motivation**] Philips is a global health technology solution and personal health product provider. Their purpose is to improve well-being and health through meaningful innovations. Historically, Philips started as a lightbulb manufacturing company that pivoted in the last years to become a leader in health technology. This was facilitated by transforming their organisation through reorganisations, mergers, and divestments to become a focused company.

[**Question/problem**] The IT landscape of Philips has drastically changed. Systems have been added, replaced, and merged into each other as the company grew and made steps to focus its IT landscape. Different processes had to follow these changes in how to manage data. Overall, this put stress on the organisation to maintain correct master data consistency and quality across each relevant system which impacted the operational efficiency of current processes. Specifically, the existence of multiple records (duplicates) of the same supplier created problems in the Purchase to Pay process with the initiate a purchase from a supplier to its final payment activity. As these activities span multiple systems, there is a clear challenge to overcome this Supplier Master Data integration problem. We formulated a main research question: ***How can Supplier Master Data be integrated into the organisation?***

This question was answered by starting with a systematic literature review to gather foundational knowledge on existing literature on data integration for Supplier Master Data. Four different integration categories were identified: 1) Governance, 2) Architectural, 3) Integration modelling, and 4) Integration algorithms. We used integration algorithms in an initial design followed by iterative improvements with the stakeholders of Philips to finalise the integration solution design. The Action Design Research methodology guided the latter process.

[Results] Our work focused on developing a Data Integration component that identified several different integrations algorithms. We found JaroWinkler to be the best suitable for the Supplier Master Data application with some optimisations for several data quality issues. This was combined with a Ranking Rules sub-component which used the algorithm's output, i.e., matched supplier records, to identify which supplier would persist (survivor) and which records would be merged (loser). This allowed the survivor supplier to be the single source of truth within the IT landscape.

The Data Integration component was preceded by the Data Sources component and followed by the Visualization & Consume component. The first served as fundamentally as the preparation of the data in terms of how it is modelled in relationships, attributes, and source systems. The second was aimed at providing the ability to validate the integration results by providing all essential information at a glance. Combined with a proposed risk-based prioritisation, we achieved an efficient and effective review to fulfil the organisation's requirements.

[**Contribution**] This research proposed an integration solution design for Supplier Master Data to maximise integration effectiveness with minimal manual review. It was implemented in practice and successfully achieved integration that resulted in a 35% decrease in the Supplier Master Database of Philips.

1 Introduction

Philips is a health technology company aimed to improve people's lives through innovation and products. They are a large global provider of healthcare products & solutions to commercial partners, such as hospitals. Additionally, their range of personal health devices promotes a healthy lifestyle for consumers.

Philips was started in the late 19th century with a specialization in lightbulbs. A business that they have since departed from, yet the association of Philips and lightbulbs remain strong. Historically, Philips was a diversified company with an extensive portfolio of consumer goods and even semiconductors products. Nevertheless, in the last two decades, they have become more focused. This led to the splits of several businesses and divestments, resulting in companies such as ASML, NXP, and Signify. Today Philips has two primary business focuses: Personal Health (PH) and Health Systems (HS). Respectively, their consumer goods and commercial health products & solutions.

This was attained through a track of R&D and, specifically for HS, through acquiring (health technology) companies over the years. Attempting to gain a cutting-edge advantage over its competitors in business. However, these acquisitions created challenges for both Philips and the acquired companies. Most relevant to us: adopting a new IT landscape and everything associated. Acquired companies continue to exist within their environment or are onboarded to the Philips IT landscape. This requires a large and complex combination of system and data migration. They are utilizing an existing system or bringing their system(s) to the larger landscape. Continuing this practice could stress managing a coherent set of master data in addition to the 'regular' challenges. This impacts all operational activities of an organization due to master data being at the core of each business process.

This thesis will explore the master data management challenges that Philips observes. In particular, as mentioned earlier, the challenges created a large and diversified IT landscape coupled with data inconsistencies. We will explore how we can support Philips in overcoming these challenges through research primarily in data integration and, to some extent, in the larger field of master data management. As this is a very real-world problem and all the associated consequences, we set out to maintain scientific rigour while providing solutions that add tangible value for Philips beyond pure research.

1.1 Problem statement [REDACTED]

1.2 Research Questions

Based on the gaps identified in the literature review and the problem statement above, we can formulate a single main research question with related sub-questions. This will allow us to frame the research scope and find a solution to the problem. A description of the research questions and their goals will follow:

Main RQ: How can Supplier Master Data be integrated for the organization?
The main research question is to address the problem statement experienced by Philips. This should include a holistic solution that addresses the problem through data integration and activities to sustain the solution specifically in the Supplier Master Data domain.

Sub-RQ1: What integration algorithms can be applied to Supplier Master Data? *To identify integration algorithms that have been applied in other data domains from literature and evaluate their applicability to Supplier Master Data.*

Sub-RQ2: How can an organization reduce manual review of integration results? *To identify activities in the solution that will reduce time-intensive manual review. This will provide pragmatic benefits as for organizations time is money.*

The answers to the research questions need to be developed into a coherent solution that addresses the problem statement. The sub-research questions allow us to break down the question into critical elements that require answering. Respectively, these sub-research questions translate to organizational values: 1) efficient, 2) effective, and 3) sustained. These values describe an ideal solution for the organization.

An overview of the problem and the organizational context has been provided in this chapter. The remaining thesis is structured as follows: Chapter 2 provides a literature review to gather relevant knowledge in this domain. Chapter 3 describes the methodology used to design an artefact. Chapter 4 goes through the design of the solution artefact. Chapter 5 describes the validation with the organization. Chapter 6 shows the updated artefact design based on the feedback from the validation. Chapters 0 and 0 contain the thesis's discussion, limitations, conclusion and future work.

2 A systematic literature review on data integration techniques for master data management

2.1 Abstract

[**Context/motivation**] Matching master data between sources/systems is an everyday activity and necessity for organizations to fulfil business activities, e.g. analytics, operational efficiency and transactions. An overview of relevant techniques needs to be included to guide this process.

[**Question/problem**] There has been a lack of systematic review of data matching techniques, specifically in master data management and their related challenges. We performed a review of research that has their application set within the context of Master Data Management (MDM), the process of managing the master data of an organization to leverage its value entirely.

[**Results**] We have reviewed 31 papers that researched data integration. Our results show that there can be different integration techniques ranging from more design-based to more immediate measurable solutions, respectively: Governance, Architectural, Modelling and Integration Algorithms. This research is predominantly applied in the People Data, general and product domains. The evaluation was observed to be usually qualitative, but quantitative was common only if there was a familiar data set.

[**Contribution**] This paper identified data integration techniques from 2007 and 2019. It provides an overview of utilized techniques and their associated challenges, data domains in which research was applied, and evaluation methods. Current gaps in the research field have been identified as well.

List of keywords: entity resolution, deduplication, data matching, data integration, master data management.

2.2 Introduction

Matching data is critical for organizations to identify a single source of truth of an entity within the organization's IT landscape. It is an activity within Master Data Management (MDM) which is purposed to integrate, analyze and exploit the value of a company's data [1]. It includes all the activities to create an integrated data set centrally governed and leveraged for further business growth through analytics [2]. pecific data quality challenges can arise as the amount of data or complexity of the IT landscape increases, such as maintaining accuracy and consistency [2]. Alongside the increase in complexity of organizations, it is imperative to have an accurate, consistent and complete view of the data. Having accurate and consistent master data (created once, used multiple times) allows the organization to make accurate analyses and maintain operational excellence [3]. Data-driven decisions are based on data analyses and are more beneficial than feeling or intuition-based decision-making [4]. From an operational perspective, the lack of master data can introduce errors and inconsistencies. For example, a delivery and invoice document will be created after a customer purchases. If the customer has different representations of itself within the data, e.g. an identical company with multiple locations, the documents might reference different addresses. Resulting in sending the product to the wrong location and the invoice not being paid out as the office location does not recognize the purchase. Correct master data would have provided a consistent view of the customer and its information across different systems. To achieve this, data has to be matched across multiple sources to create master data and allow for MDM.

MDM is purposed to integrate, analyze and exploit the value of the data of a company [1]. The process can be described as an iterative cycle with the following activities [5]:

1) Identify key organizational data objects;
2) Semantic harmonization;
3) Data integration (data matching);
4) Enrichment of initially integrated master data;
5) Monitoring data quality.

In step 3, data integration is a critical activity consisting of matching, normalising, cleansing and synchronising master data from different sources. It identifies which record pairs across databases are a single entity. Matching data without optimisation across two databases would require comparing all record pairs for an m x n number of calculations. Master data can be characterised by low change frequency and constant data volume and typically consists of entities related to customers, products, employees, or suppliers (vendors and/or manufacturers) [3], [5]. However, with the increasing volume of data, any changes in the source database can potentially significantly impact the computational complexity. Additionally, the IT landscape of an organisation can consist of multiple systems and databases, requiring the integration process to occur multiple times. Computational complexity is directly related to the amount of time and memory required to compute it poses challenges to the MDM process to sustain a fast, effective, iterative management of master data.

Data matching research can find its roots within the healthcare domain, where researchers provided a probabilistic method to match records [6]. They introduced a methodology that integrated indexing, weight, and classification. Indexing provides calculation enhancements

as it is usually only feasible to compare some record pairs within a reasonable timespan when the datasets become very large. Indexing would result in a subset for calculations called a block. Weights interpret the likeliness of a record pair belonging to the same entity based on a field or entire record comparison algorithm. These algorithms can range from simple exact string comparisons to more complex algorithms. The comparison algorithm can output different types of weight. This could be a value ranging from zero to one where one is an exact match (probabilistic) or a 'match' vs 'non-match' result (deterministic). A classification provides the range of weight considered a match, no match, or anything in between. Ultimately, the classification is the end-state determining whether a record pair reflects accurately to a single entity.

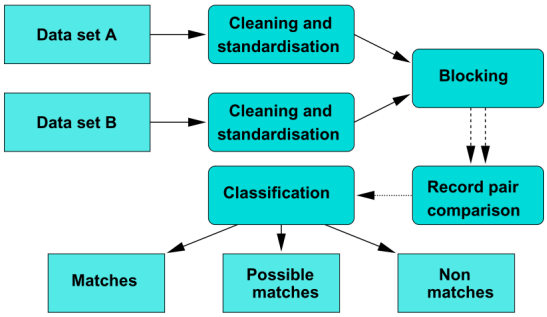

Figure 1 Generic steps for matching data across two data sets.

There have been many studies that explored 7 data matching techniques. However, few have focused on applying data-matching techniques within an MDM process. As a result, there is a gap in how to account for the following challenges:

- Accounting for computational complexity due to the volume of an organization's data.
- Accounting for dynamic environments, i.e., changing production system(s) data.
- Prioritization of data entities (with high business impact) to be matched.
- Application/empirical testing on master data sets (e.g. customer or supplier master data).
- Most methodologies showcase the ability to match (static) datasets against each other. Firstly, there are constant creation, update and deletion operations on the records of a database. Even though master data remains relatively stable (compared to transactional data), it should account for any changes that happen when data are matched against each other [5]. Secondly, a prioritisation method is needed to identify the most critical data. Especially when the datasets are significant and computational time increases, a company can consider focusing on its most essential records followed by decreasingly important data and, for example, focusing on your top 100 customers that account for most of your business. Lastly, methodologies performed on different data types than master data would be irrelevant. For example, comparison algorithms that work well on full texts would not be comparable to master data as the first would focus on semantics and the latter on its value. There is a need for a methodology that allows for incremental matching of master data to account for changes in the data and prioritisation of records. This aligns with MDM as it is a continuous process, not a one-time exercise. This study aims to identify literature that researched data matching techniques applied within an MDM context and/or have been applied to master data.

2.2.1 Review goals and research questions

The systematic literature review has three goals: (1) to provide an overview of data matching techniques, (2) to create an overview of evidence for data integration techniques that have been applied to master data and (3) to identify challenges that researchers have identified with matching master data.

RQ1: What techniques have been researched for data integration in MDM?

RQ2: What are identified challenges for data matching and integration in literature?

This literature review will contribute by providing an overview of studied data matching techniques and those that have been applied in an MDM context and/or have been applied to master data. It is expected to identify solutions, additional challenges, and learnings.

2.3 Key Concepts

2.3.1 Master data management and master data

Master data management has various definitions, but we choose to use the following definition: all activities for an organization to manage its data and to leverage it to integrate, analyze and exploit the value of its data assets [5]. It encompasses the goal of master data management without trying to scope its activities exactly. Any activity that serves this goal would be relevant in this area. It is an activity critical for every modern (data-driven) organization that utilizes various tools to support its processes. For example, master data is a fundamental building block for vendor purchasing processes. It provides core information about who the vendor is, where it is located, and how to contact them (i.e. master data), and this is constantly referenced in further transactions such as purchases from them (i.e. transactional data). Therefore, having correct master data is essential as it can have a cascading effect due to being referred to so often.

2.3.2 Data matching techniques

Research on data matching has resulted in many methodologies for data matching [7]–[12]. Indexing evolved from traditional blocking, fixed range for all comparisons, towards techniques incorporating clustering or machine learning approaches [13]. Field or record comparison algorithms research has seen many variations on name-matching algorithms, e.g. (edit) distance/similarity or abstraction/phonetic [9], [14]–[16]. Classification research range from rule-based and probabilistic approaches towards learning algorithms that can determine matching criteria with or without supervision (machine learning) [17]. Studies have presented methodologies that guide on data matching from start to finish. However, these methodologies are not comparable as the utilized techniques, preparation and origin of the dataset(s) are different from each other and do not have a wide range of techniques in scope [18], [19].

2.3.3 Data matching in MDM

Several studies have investigated the role of data matching in an MDM context. Entity Identify Information Management (EIIM) incorporates the complete identify management of an entity [20], [21]. These studies introduce a record-based and attribute-based mapping technique. This attempts to match entities based on the (exact) values of the records or attributes. Therefore, (minor) data inconsistencies can result in missed matches. Such an approach can be practical if data is consistent and standardized across the to-be-matched databases. This is a similar approach to instance-based resolution techniques introduced in a different study [22]. These studies do not detail a data matching technique that includes an indexing technique, matching algorithm or classification to enable faster computing and account for slight deviations in string-type attributes to match the records to a single entity. As timely data-driven decision-making is crucial to react to competitors to gain or sustain a competitive advantage, efficient data matching must be achieved through indexing [23]. his allows the matching algorithm to narrow down the records that are most likely to be matched. All companies face data quality issues such as data consistency. Therefore, it is crucial to account for these inconsistencies with a comparison algorithm that provides a score and a

classification that quantifies the likelihood of the records belonging to the same entity. High data quality is a pre-condition to leverage value from your data [4], [24].

2.4 Research methodology

Kitchenham's Systematic Literature Review guidelines will be followed [25]. A systematic review protocol is essential for stating clear objectives, search method, assessment of the validity of the findings and an unbiased systematic presentation of findings [25], [26]. A systematic literature review would provide the highest inclusion of data matching techniques relevant or have been applied within an MDM context.

2.4.1 Search strategy

Both scientific and grey databases will be searched. The scientific databases will yield peer-reviewed literature, and they would have a bias towards positive results of the data matching technique. Grey databases are included to account for publication bias and to get the most recent data-matching techniques developed and/or performed on real-world data. The grey databases are expected to yield literature that has been applied from a practical perspective first instead of a scientific perspective. Particularly for organizations, practicality and good performance can outweigh scientifically sound methods. The databases have also been selected based on a systematic literature review that focused on the distribution of MDM literature across years and databases [27]. This allows us to include the most promising databases that includes MDM literature.

Scientific databases:

- SCOPUS (https://www.scopus.com/home.uri)
- ScienceDirect (https://www.sciencedirect.com/)
- ACM Digital Library (https://dl.acm.org)
- IEEE (https://ieeexplore.ieee.org/)

Grey Literature:

- Google Scholar (https://www.scholar.google.com/)
- University of Twente Repository of Student Theses (https://essay.utwente.nl/)

2.4.1.1 Initial search in Scopus

An initial exploration phase will be performed to test the setup of the systematic literature review:

1. The search strings on the number of studies it returns will be tested. An overview of the resulting studies based on the search string will be provided.
2. Studies will be sampled for data extraction to check if any additional data needs to be extracted and if the results are relevant.
3. The studies will be assessed on quality and if they align with the research questions of this review.

Initially, the scope was set on providing an overview of all data matching techniques, which included indexing, record comparison algorithms and classification techniques. However, this resulted in several thousand search string results in the Scopus database. Reviewing each study will only be feasible using sophisticated text mining solutions as these techniques were not always clearly defined in the title nor elsewhere in the papers. Alternative analyses can be considered. For example, analyzing from a geographical or authoring perspective, grouping by type of technique studied, distribution of publication source (scientifically or grey – potentially more practical research) or data matching technique.

The search string needs to include additional criteria to narrow the search results for RQ1 and RQ2. The addition of "master data management" and synonyms resulted in 17 identified papers in Scopus. Experimenting with different synonyms did not yield favourable results. Therefore, expanding the data matching synonyms provided 86 results. For RQ2, additional criteria with "challenges" or "learnings" narrowed the results to 21. However, these results are expected to be a subset of the previous string, and therefore, these additions were kept out.

Due to the number of resulting papers, no limitation has been set on the year of publication, initially, from 2012 onwards, as there was already a collection made of several data matching techniques [7]. Unfortunately, this collection is a published book and did not provide any scientific methodology.

Search string exploration for an overview of data matching techniques:

- ("data matching" OR "data linkage" OR "entity resolution" OR "object identification" or "field matching")

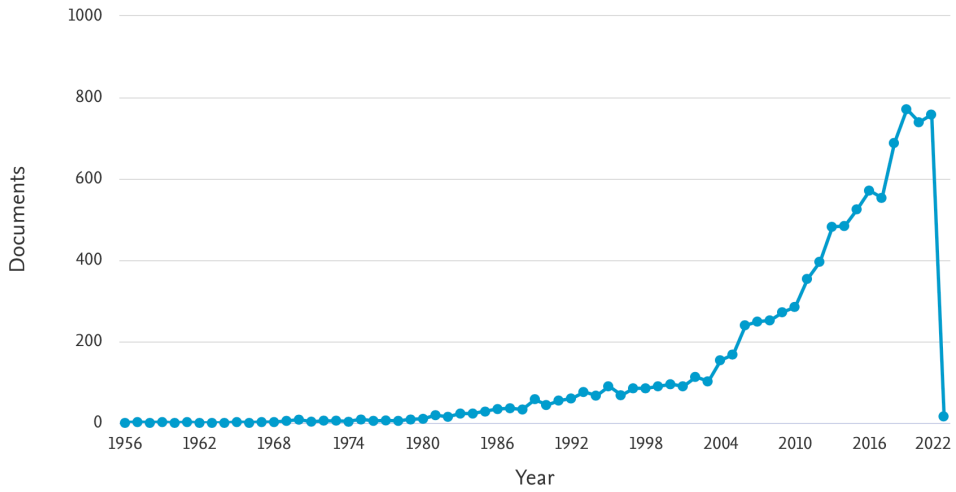

Figure 2 Scopus – 9000+ results

Search string for RQ1 and RQ2:

- (“master data management” or “mdm” or “master data”) AND (“data matching” OR “data linkage” OR “data integration” OR “entity resolution” or “record linkage” OR “deduplication”)

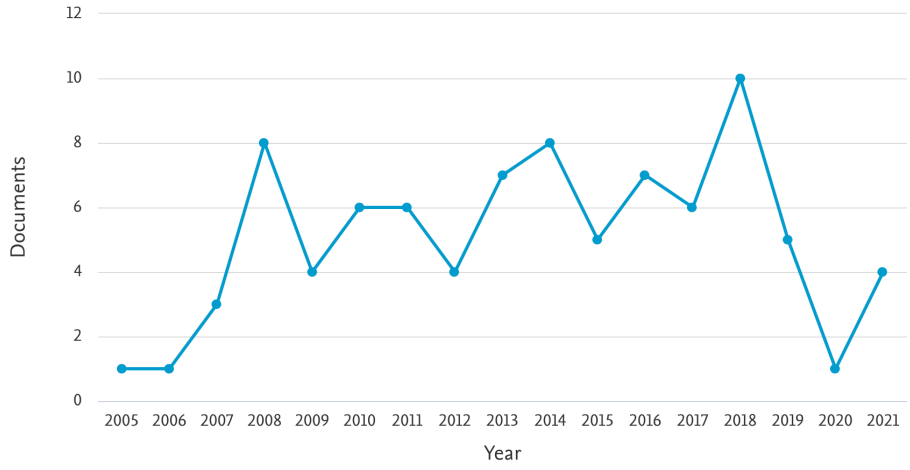

Figure 3 Scopus - 86 results for RQ1 and RQ2

2.4.1.2 Expanding the search to other databases

Applying the search string to other databases (ScienceDirect, ACM Digital Library, IEEE, Google Scholar) yielded a considerable variation in the number of papers found per database. Potentially due to the different search algorithms and/or the filtering settings. For example, the definition of subject areas or types of documents. Subject areas were only sometimes consistent across the databases, but they were crucial in filtering down the results. Also, document type definitions differed between the databases: “research article” vs “conference paper” vs “article”. By common sense, standard filtering settings were applied; however, the massive variation of results showed that the reproducibility of the initial search string was low.

Scopus	ScienceDirect	ScienceDirect (Title-abs-key)	ACM Digital Library	IEEE	Google Scholar
86	124	4	192	6	142
Limit to relevant subject areas "Computer Science" "Business, Management, and accounting" Limit to "Articles" and "Conference Papers"	Narrow down on research articles as they are structured scientifically relevant (reviewed, research questions, conclusions, methodology)		Narrow down on research articles as they are structured scientifically relevant (reviewed, research questions, conclusions, methodology) PDFs only		
60	15		124		
	Limit to "Computer Science" and "Business, Management, and Accounting"				
	11				

Figure 4 Breakdown of the filter steps and the number of results

Due to the inability to achieve consistent filter parameters across the different databases, an alternative database was explored. FindUT provides a search through multiple databases, allows the search parameters to be consistent, and, therefore, should provide a consistent

number of results for the literature review. Unfortunately, after sampling the search results, the papers were mainly deemed irrelevant to the search string.

2.4.2 Study selection

The following criteria should be met to determine if an article classifies as a primary study:

Inclusion criteria:

1. The study is written in English
2. The study contributes to one or more research questions
3. The study is with the retrieved with the defined search string in the title, keywords or abstract
4. The study is limited to subject areas Computer Science or Business Management and Accounting

Exclusion criteria:

1. The study does not meet the inclusion criteria
2. The study should be retrievable in full-text
3. The study should be retrievable in a PDF format
4. Studies in non-standard form (e.g. posters, presentations, web articles)

The initial study selection criteria resulted in only 19 papers as relevant by reading through the title, abstract, introduction and, if necessary, any other chapters of the paper. Further analysis showed that the criteria needed to be narrower on data matching techniques introduced under This literature review will contribute by providing an overview of studied data matching techniques and those that have been applied in an MDM context and/or have been applied to master data. It is expected to identify solutions, additional challenges, and learnings.

Key Concepts. Therefore, papers focusing on data matching techniques such as integration on a database level, utilization of schemas/mappings, utilization of ontologies and metadata, and through an architectural setup were excluded. These papers resulted in 34 papers for further quality assessment.

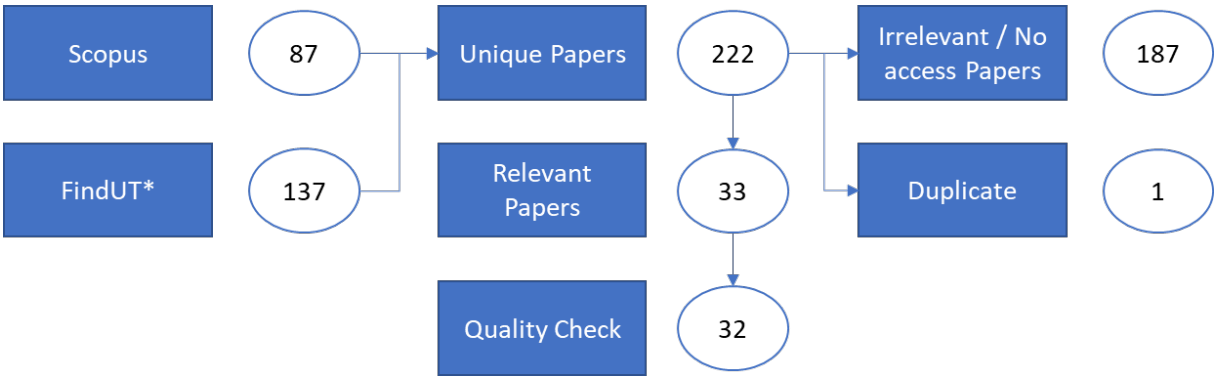

Figure 5 Study selection process

2.4.2.1 Analysing FindUT search engine

All FindUT results were deemed irrelevant through our inclusion/exclusion criteria. This is a surprising finding and requires additional investigation. The final search string utilized in FindUT was created by utilizing the "Advanced Search" option and manually including the AND/OR operators to create the search string. This resulted in the following generated search string:

- kw:(data matching) OR kw:(entity resolution) OR kw:(data integration) OR kw:(data linkage) OR kw:(record linkage) OR kw:(deduplication) AND kw:(master data management) OR kw:(master data) OR kw:(mdm)

By testing single keyword "data matching" the "Best Match" results all seem relevant at first sight. The number of results is very high 496,527. By including an additional AND operator for "master data management" resulted only 16,603 hits, top results were relevant.

Seemingly the grouping of keywords via the advanced search does not explicitly process that the keywords need to be exact. "kw(data matching)" searches for the keywords "data" and "matching" in a paper and not "data matching". This is a confusing implementation as both words are encapsulated by the function for keywords. Expanding keywords result in "kw(A, B) AND kw(C, D)". Which result in that the paper needs to contain any of the keywords A, B, C, or D. Proof of this finding with the search string "kw(data matching) AND kw(record linkage)". Example in Figure 6 does not contain the keyword "matching" within its text.

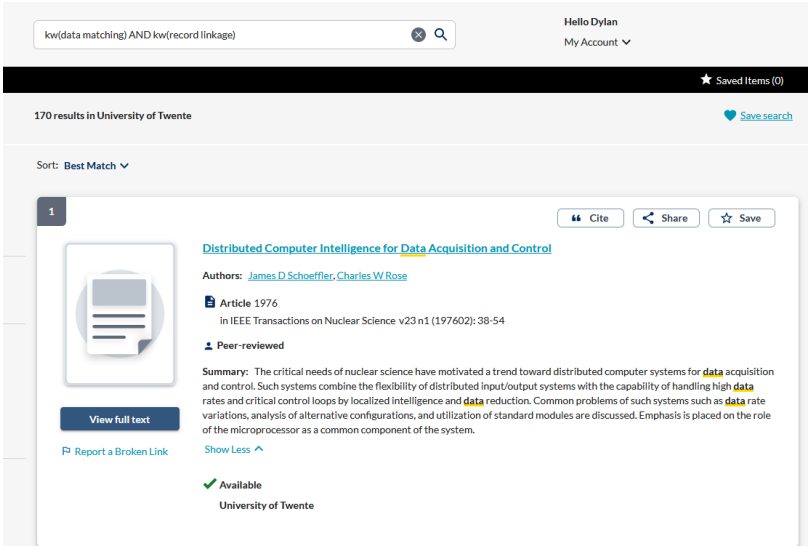

Figure 6 Example paper not matching with search keywords

When limiting the search results to "Full Text" reduced the number of results as expected. However, when also limiting to results that are "Open Access" increased the number of results higher than before. This is a strange behavior and currently without explanation.

- Search string: kw:("data matching") OR kw:("entity resolution") OR kw:("data integration") OR kw:("data linkage") OR kw:("record linkage") OR kw:("deduplication) AND kw:("master data management") OR kw:("master data") OR kw:("mdm")
    1. Articles and English only – 23,475 results
        - Filter on "Full Text" – 22,130
        - Filter on "Open Access" – 2,473,527 results
        - Filter on "Full Text" and "Open Access" – 2,473,578 results

Cilliers, Frans; Flotman, Aden Paul	The psychological well-being manifesting among master's students in industrial and organisational psychology
Burnham, S. C.; Coloma, P. M.; Li, Q. X.; Collins, S.; Savage, G.; Laws, S.; Doecke, J.; Maruff, P.; Marti...	Application of the NIA-AA Research Framework: Towards a Biological Definition of Alzheimer's Disease Using Cerebrospinal Fluid Biomarkers in the AIBL Study
Govender, Indiran; Mabuza, Langalibalile H.; Ogunbanjo, Gboyega A.; Mash, Bob	African primary care research: Performing surveys using questionnaires
Al Jabbari, Youssef S.; Tsakiridis, Peter; Eliades, George; Al-Hadlaq, Solaiman M.; Zinelis, Spiros	Assessment of geometrical characteristics of dental endodontic micro-instruments utilizing X-ray micro computed tomography
Moreira, Caio Rosas; Oliveira, Daniel Vicentini De; Vieira, Lenamar Fiorese	Rev Bras Cineantropom Hum regulation of master swimmers during the
Herbster, Adolfo Fernandes; Romero, Murilo Araújo	EDFA design and analysis for WDM optical systems based on modal multiplexing
Diacon, A.; Bell, J.	Investigating the recording and accuracy of fluid balance monitoring in critically ill patients
Akuru, Udochukwu B.; Okoro, Ogbonnaya I.	Renewable energy investment in Nigeria: A review of the renewable energy master plan
Peter, Laura; Rüst, Christoph Alexander; Knechtle, Beat; Rosemann, Thomas; Lepers, Romuald	Sex differences in 24-hour ultra-marathon performance - A retrospective data analysis from 1977 to 2012
Jung, Jisun; Lee, Soo Jeung	Exploring the factors of pursuing a master's degree in South Korea
Lira, Silvia Oliveira Ribeiro; Sousa, Vanessa Patrícia Soares de; Medeiros, Caroline Nayane Alves; Viana...	Impact of lumbopelvic pain on postural balance during sit-to-stand activity in pregnant women: a cross-sectional study
Rudolph, Heike; Salmen, Harald; Moldan, Matthias; Kuhn, Katharina; Sichwardt, Viktor; Wöstmann, Ber...	Accuracy of intraoral and extraoral digital data acquisition for dental restorations
He, Lili; Wang, Jin; Bai, Hongtao; Jiang, Yu; Li, Tonglin	A novel parallel and distributed magnetotelluric inversion algorithm on multi-threads workloads cluster
Makombe, Simon D.; Hochgesang, Mindy; Jahn, Andreas; Tweya, Hannock; Hedt, Bethany; Chuka, ...	Assessing the quality of data aggregated by antiretroviral treatment clinics in Malawi

Figure 7 Snippet of the 137 results from FindUT

2.4.3 Quality assessment

For quality assessment of the identified primary studies, several questions from Kitchenham's checklist for qualitative studies have been adopted [25]. The assessment will assist in the primary study selection and the literature analysis. It will indicate the validity of the research and environment in which the data matching technique was researched, i.e. has it been applied within an MDM context, and has the data source been described?

These questions will be scored at three values: Yes = 1, Partly = 0.5. No = 0.

- Has the data sources/attributed (characteristics) described?
    - Yes: there was explicit details on the characteristics of the data for which the data matching technique was prescribed to.
    - Partly: there was an indication of the type of data subject for which the data matching technique was prescribed to.
    - No: there was no mention of the type of data or its characteristics for which the data matching technique was prescribed to.
- Have different perspectives and contexts been explored?
    - Yes: there was an explicit explanation on various data matching perspectives within the MDM study.
    - Partly: there was an indication or reference to other data matching perspectives within the MDM study.
    - No: there was no mention of other data matching perspectives within the MDM study.
- Has the paper provided an evaluation or validation of the proposed method/artefact?
    - Yes: there was an evaluation or validation that includes either synthetic or real-world data/use cases
    - No: there was no evaluation or validation

The quality assessment resulted in 31 of the 32 relevant papers passed. The excluded papers did not provide relevant information that could contribute to the research questions. The other papers were accepted for further data extraction and synthesis. A threshold has been set on a score of 1 or higher as a paper can contribute to one or both research questions if one of the quality assessment questions is met. See Table 8 Quality assessment results.

2.4.4 Data extraction

Standard information will be extracted from the identified primary studies, see Table 1 Data collection form. Several additional data points will be extracted related to the research questions.

Extracted data	Description	Type
DOI	Unique identification of the study	Standard
Title	The title of the study	Standard
Author(s)	The author(s) of the study	Standard

Document type		E.g. article, journal paper, conference paper	Standard
Source of publication		E.g. in which journal the study was published	Standard
Data of publication		Date of the publication within the source	Standard
Keywords		The listed keywords of the study	Standard
Abstract		The abstract of the study	Standard
Goal		The goal of the study	Standard
Data Integration Technique	Indexing techniques	The technique used to optimize data matching performance	RQ1
	Matching technique	The technique used to calculate weights (similarity) between two records	RQ1
	Classification techniques	The classification technique to determine the threshold when two records are to be considered a single entity	RQ1
	Other techniques	Any technique that does not fall under indexing/matching/classification techniques	RQ1
Data domain		Data domain in which the data matching technique was researched in (e.g. customer master data). Properties, volume, origin of the data	RQ1
Data matching evidence		All evidence related to the data matching methodology and results.	RQ1
Identified data matching challenges		Challenges that have been identified within the study related to a single or combination of indexing, matching, classification techniques or MDM. This can include input from the future work section	RQ2
Identified data matching benefits		Benefits that have been identified related to the utilization of a single or a combination of indexing, matching, classification techniques or MDM	RQ2

Table 1 Data collection form

2.4.5 Data synthesis strategy

The data synthesis strategy is expected to be qualitative. Based on the research questions, the goal is to identify the researched techniques, empirical evidence, challenges and benefits learned. These occur in the sections of the results, conclusion, and future work/contributions. To identify if the study has been applied in an MDM context, the research methodology might also be a viable source of information. Results will be presented in the structure of the research questions.

Quantitative data might be available in some studies, such as the performance of data matching techniques, e.g. F-score (precision and recall). However, these would be incomparable as studies will have set their parameters in the data matching algorithm, utilising different data sets and using their measurement metrics. Therefore, any quantitative data will be excluded from the results.

2.5 Results

This section presents the findings for the research questions. It will provide the results for **Sub-RQ1 (What integration algorithms can be applied to Supplier Master Data?)** with an overview of all the identified data integration techniques, data domains and evaluation techniques. The results of **Sub-RQ2** (**How can an organization reduce manual review of integration results?)** will be presented right after each data integration technique/research, as the challenges and benefits are strongly dependent on the technique.

2.5.1 Integration Techniques

The following integration techniques have been described and researched in the included papers. The techniques have been categorized into four themes that focus on different areas to achieve data integration: Governance, Architectural, Integration modelling, and Integration algorithms. This shows a progression of techniques that impact integration at a high-level (e.g. roles and responsibilities) to more low-level techniques that influence the data directly (e.g. record pair scoring algorithms).

2.5.1.1 Governance

Governance describes the roles, responsibilities, processes, and standards to maintain data within an organization. It defines who should perform what activities to ensure data is maintained at a level that enables effective utilization. This concept is not a direct integration technique. However, it is an enabler to have control of your data ownership and quality before you proceed with integration steps. Two studies have been found that describe how governance can be structured.

2.5.1.1.1 Roles and responsibilities

Two studies recommend a structure where data is owned and maintained by their data owners [28], [29]. These owners hold the right of decision-making and are responsible for organising the assets. This includes defining a data quality standard, execution of improving data quality improvements and ensuring data assets are made available.

[29] differentiates between IT Governance and Data Governance. Whereas the latter only focuses on managing data quality to support business goals. This shows a split between an organisation responsible for defining how data should be organised (e.g., policies and standards) and an organisation responsible for maintaining the data to the defined policies.

[28] prescribes a centralised approach where a group of data assets or systems are owned by a single owner or are overseen by a common ownership entity versus a decentralised approach where individual data assets have various owners.

2.5.1.1.2 Generic approaches

One paper describes several data integration approaches to achieve an integrated view of the data at the highest level [30]. Firstly, manual integration, where data is collected, cleaned and matched by hand between two or multiple systems. Secondly, using middleware or applications enables a connection between data sources with constant data transfers. Thirdly, uniform access integration by connecting all data sources into a standard view. However, this focuses on data representation and does not impact the source of data itself. Lastly, standard storage integration extends the uniform access integration but stores a copy of the unified view.

Three key challenges related to the volume, structure and understanding of data are identified:

1. Large volumes of data can result in accessibility issues due to its size or the spread of systems and their stores. This also means that various teams own/maintain these sets with different configurations.
2. The data structure can hinder integration due to logical data model incompatibility, duplicates of values or differences in the data type, structures and unstructured.
3. (Domain) understanding the data is vital in leveraging the data to support business activities.

Achieving data integration with a business context or goal is meaningful. Identifying the correct business logic to support the integration is necessary. However, identifying the right people that understand the various data sources and aligning them is a challenge that lies beyond the capabilities of conventional data integration methods.

2.5.1.1.3 Challenges

Technical readiness and incentives among individual data owners may result in a siloed and ineffective approach to managing data [28]. This leads to desynchronization of how data is maintained (stored, data descriptions, adherence to standards) and inconsistent availability. Different technologies can be used to access data with different authorizations if data owners are not on the exact technical implementation and readiness to collaborate.

2.5.1.2 Architectural

Architectural integration techniques describe concepts, the design of systems and (communication) layers to enable flows of data within or between systems. This could include a design to enable Extract-Transform-Load (ETL) operations to prepare and cleanse data, utilization of other data sources for enrichment (e.g. metadata), and processing for sending and storing data from source to target location(s). Additionally, this could include a design of how data is managed (e.g. in a centralized manner).

2.5.1.2.1 Standard communication protocol between data sources

The first architectural method was observed to be a standardized communication protocol between data sources. An architectural design that describes the connection between components such as middleware and service-oriented modules to enable connections

between different data sources [31]–[37]. The main focus is to ensure that all components communicate with each other in a standardized manner. A concrete protocol implementation was proposed by utilizing XML schemas [38].

2.5.1.2.2 Real-time processing

The second architectural method was real-time processing. Several papers propose a method to integrate data with a high volume of data and/or the requirement to have time-critical integration activities. Therefore, (near) real-time processing must be included in the architectural design to enable large amounts and fast computing. This was achieved through batch-wise and streaming of data within memory supported with specialized hardware and optimized joining algorithms/operations [39], [40].

2.5.1.2.3 Challenges

Identifying and adapting to data schema changes across multiple data sources is a challenge. Each data source would require some monitoring capability to identify changes and follow up with the adaptation to the changes. This would require time, resources and time investment. Therefore, adhering to a standard communication protocol would remove the necessity to track these changes. Communication can occur as long as the different sources utilize the same protocol.

Additionally, the exponential growth of data processing and storage requirements creates a clear challenge to enable timely data availability. Performing data joins on enormous datasets is a very time- consuming operation. In domains where the timeliness of data is critical for business decision- making, this creates a need for fast and efficient joining methods. Additionally, the speed is highly impacted by the quality of the hardware where the data is stored and processed. Traditional hard discs are read mechanically and would benefit from large in-memory storage. Therefore, architectural decisions need to account for these challenges and include some future-proofing.

2.5.1.3 Data modelling

Modelling for integration describes the artefacts that support the convergence of heterogeneous data sources. This could be at a data level in the form of schemas, data models, structures, or this could also be at a descriptive level in the form of metadata, semantic descriptions, or an ontology. These modelling artefacts provide information to integrate data sources.

2.5.1.3.1 Harmonization of data through various methods

Methods to converge between different data sources was research in various ways: data modelling [41]–[43], annotations (semantic and ontology) [44], [45], and centralized management [46]. The first overcomes differences in data structures by defining a common structure to follow or to design the data model to be compatible with each other. The second focuses on providing annotations so that different data sources can align by following a semantic network or ontology. The last combines the storage of different data models in a central repository. Supported with a description of how each data model in its source was utilized (context).

2.5.1.3.2 Challenges

Traditional data model integration techniques involve pre-defining schematics to overcome differences between data sources. This can be a time-consuming activity, is increasingly complicated and reduces flexibility when the number of sources increases. Data deduplication and integration depend on the availability of a single source of truth across different data sets or systems. This includes overcoming differences between data models and inconsistencies in values and storage formats. These differences must be resolved to identify a single source of truth.

2.5.1.4 Integration algorithms

Integration algorithms can fulfil a combination of indexing (blocking), matching, and classification purposes. Indexing provides an optimization of compute by eliminating those records that are most likely not to match up to each other. Matching or record-pair scoring is the calculation of how similar the records are. Classification provides the threshold of the matching to determine whether the records match each other.

There were four main integration algorithms observed: edit-distance comparison, token-based comparison, supervised and unsupervised Machine Learning. One paper identified a file-integration based method which was standalone by itself. The algorithms are further elaborated in Section 4.2. An overview of the papers and its researched algorithm can be found in Table 9 Overview of research and integration algorithm.

2.5.1.4.1 Challenges

Quality is critical in optimizing precision and recall. This reflects on the availability and completeness of data. The scalability of the method to larger sets of data can exponentially increase the integration and training times of the model. Optimizing calculations with the use of effective blocking functions are critical. Speed of development of the system is important to adjust if necessary to improve precision and recall. This ability allows for rapid development of variations of the system and selection of the top performing one.

The ability to match data irrespective of the data source is a challenge. It traditionally requires human efforts and a dependency on (clean) training data. Therefore, matching data via the traditional way requires knowledge about the data and time-consuming efforts to provide the necessary inputs to have accurate matched data. A method that allows for automatic identification of the quality of data and what the (expected) correct matching patterns are will resolve this challenge.

2.5.2 Data Domains

The following data domains have been described and researched in the included research papers. These data domains were either specifically described as the intended use case and/or have been used as an example of the data integration research. Several data domains have been covered and a categorization has been made based on the characteristics of the specific domain. See Table 10 Data Domains.

2.5.2.1 Master Data

Several papers applied their research on the master data domain. Where this specific domain is characterized by having a core set of data created once (a single version of truth) and used multiple times. These references can occur multiple times within and/or across systems. This is a very generic description of its purpose of master data and does not include specific requirements on the characteristics of data, e.g. value constraints, inclusion of certain fields, or relationships. Therefore, it can be seen as fundamental reasons to this domain and other data domains for matching to achieve certain key objectives:

- Ability to define a core set of data and its properties to use and reference across multiple data sets/systems, e.g. the data model, value and relationships constraints, naming convention, purpose of use.
- Ability to create a unified view of all the data irrespective of the source(s).
- Efficient data management by maintaining data centrally that is referenced multiple times.
- Effective data management by controlling a core set of data to certain quality standards and ensuring purpose of use is achieved, e.g. accurate analyses, unique reference in transactions.

2.5.2.1.1 Aviation

For the Aviation sub-domain, there is the need to provide transparency for auditing and controls purposes. Therefore, the process and results of mapping data to each other needs to be logged. The paper achieves this by accounting versioning of the data both locally and the resulting master data. With this approach, they can review and compare the results of the data integration and track exactly what local version was used, how the data integration steps were determined and the master data created.

2.5.2.1.2 Product

The Product sub-domain is characterized by a stronger dependency on a combination of the name of the product with its purpose of use and additionally there is the availability of global/local identifiers for some products. Identifying correct product data is critical for supply chains. Ensuring the same product is referenced across suppliers can be determined by name or identifier. Matching on product names is an obvious link, but this can result in a match between two products named identically whilst being very different, e.g. both named ‘hammer’ but one can be a ‘sledge hammer’ and the other a ‘nail hammer’. Adequate data quality, metadata and contextual information is critical accurately identify the same product. Even more complexity is created if brand names are accounted for. Brands can opt for specific naming conventions as part of their marketing strategy.

To provide some international standardization, there is the availability of global references for products that capture description, price, size and more. There is the possibility that a product is included one or multiple of these databases, e.g. Global Trade Item Number (GTIN), European Article Number (EAN) and Unique Product Code (UPC). However, they are maintained by different organizations, scope of specific product categories and are dependent

on correct usage. Nevertheless, this allows for additional cross-referencing to support matching products to each other.

2.5.2.1.3 Geolocation

The Geolocation sub-domain is characterized by the decentralization of governance and wide variance in the time origin of the data. The first characteristic is observed as geolocation data sources stored in various systems, originating from different countries with their own legislation and owned by various organizations. This lack of overarching governance puts a strong demand on individual effort to manage the data. Therefore, increasing the risk of divergence versus a unified approach on managing data. This creates a scenario where integration of data is dependent on technical readiness of the sources, i.e. adapting legacy systems to be compatible with modern systems, and willingness of the owners to participate in data integrations. Incentivizing these owners and aiming to centrally manage (with standards) will be key activities to be performed alongside the expected integration steps.

2.5.2.1.4 Unstructured Data

The Unstructured Data sub-domain is characterized by the inclusion of unstructured data to support integration of master data. Unstructured data can be described as data where there is no clear relationship between entities such as plain text, reports and e-mails. This creates challenges to accurately identify relationships between data as there is no identifier to rely on, it can be in various stored file formats that need to be processed in their own respective way, and this creates the necessity to rely on the narrative of the unstructured data. In particular with the last challenges, the narrative (i.e. meaning) can only be derived by using sophisticated tools such as Natural Language Processing to identify what the unstructured data is about and which others are similar.

2.5.2.2 Big Data

The Big Data domain is characterized by the integration of data where there are special demands to account for the sheer volume and/or the velocity of data. The volume of big data can be described as extremely large data sets where operations on it can only be achieved with specialized hardware and software and the velocity of data where the amount of new or changed data is very high with potentially a continuous stream of incoming data. Therefore, requiring sophisticated integration techniques that need to be highly efficient in computational complexity and preferably run in a distributed fashion, i.e. on a cluster. The design of the integration techniques should find an optimal balance between required time and results through blocking techniques or integrating incrementally/batch-wise on the data. In combination with the ability to perform the integration activities in a distributed fashion, where the computation is executed parallel instead of sequentially, can drastically reduce the necessary time. The key difference with parallel computation is to account for all possible entity matches and provide them to the same node. Essentially, ensuring that a comparison is made with all potential records. This could be accounted for in a pre-processing step and clustering the similar data at a high-level.

For any Big Data implementation, there is significant room for improvement by tuning various parameters such as memory usage, space allocation (in memory vs on-disk), cost model

utilization, hardware and optimizing queries. These general tuning areas would also be relevant for the purpose of data integration.

2.5.2.3 People Data

The People domain is characterized by the data that includes information about individuals through Personally Identifiable Information (PII). However, depending on the purpose of use there can be distinct variations within this domain. This can result in different steps for integration purposes and also for compliance/legal reasons. With the potential use of existing databases to reference or requirements and consequences in how people data is processed.

2.5.2.3.1 Customers

The first sub-domain of people can be considered customers. In this context it is referring to individual who act as a generic consumer, e.g. purchasing an item from a vendor. This specific sub-domain of data is commonly utilized in Customer Relationship Management (CRM) systems where the goal is to maintain a relationship with the customer. Communication is key in this context enabling customer relationship managers to assess current and future needs based on historical patterns.

It is industry-standard is to have them customer accounts with the vendor of choice with their PII. Usually unique identifiers are not required, i.e. (governmental) social security number, but name, address and e-mail are mandatory. This could provide leading factors for integration to match customers across vendors or CRMs. As generally one e-mail account is (re)-used, it is unusual to misspell your own name and an address should give you a unique geographical location. However, the challenges are that customers can have used multiple/different e-mails, the exact value mapping of a name and address can change such as having an address as a single string versus split into city + street + house number, address standardization can vary and is prone to spelling errors. Tailored integration techniques for customers is a must to account for these characterizations.

2.5.2.3.2 Employees

The second sub-domain of people is employees. This sub-domain is characterized by (mandatory) inclusion of sensitive PII such as social security number, bank account, and health insurance number. Usually required for important governmental services or for HR purposes of the employer such as payment. Therefore, making integration between sources of this purpose simple. However, some integration characteristics might present themselves in the form of value standardization. Identifying numbers could be formatted to a certain minimum or maximum length resulting in leading zeros or split data. Accounting for this is straightforward.

2.5.2.3.3 General individuals and public records

The third sub-domain of people is general individuals and public records. This specific sub-domain is defined due to the lack of uniquely identifying numbers such as the social security number to point conclusively towards an individual. For example, the same social security number between two systems point to a John Doe and a Jane Doe. These are still susceptible for human input errors. This domain is characterized by the absence of standardization of

capturing name, address and optionally more, specifically for purposes where the criticality and consequences are less as compared to the domain of employees and where there might not be a reference to a central governmental database to pull the individual's information from. These could all include records of an individual that are separately maintained and at risk for negligence. Integration techniques will have to account for different data standards, e.g. name in full, abbreviated or first letter only, and potentially also the origin of the data source. A military database with John Doe could point to a different individual compared to a municipality database.

2.5.2.3.4 Medical

The fourth sub-domain of people is medical records. This specific sub-domain is defined due to a specific situation where the availability of a unique personal identifier is present however due to legal, ethical and privacy reasons not that straightforward to use. There is overlap with the employee's sub-domain where social security numbers (or an equivalent) can be used as way to link data reliably. Alternatively, characteristics mentioned in the general individuals' sub-domain can be leveraged with considerable risk of the consequences in case the integration is faulty due to the medical nature. Therefore, these integration techniques should always include a confirmation by the relevant individual for verification.

Specific complications arise due to two identified purposes to integrate medical data: to create a complete medical view or to perform medical research. In both cases, common master data characteristics apply and care of processing data during integration and consent need to be provided to grant permission to use individuals' data. Specifically, in the second case, this can become a cumbersome activity where research is attempted on a large population and accompanied with additional requirements due to it being a medical domain. This requires inclusion of the necessary steps to account for ethical, legal and privacy reasons. This demands a more vigorous data management protocol where everything is documented, transparent for review, and approvals are gained before further steps are taken.

2.5.3 Evaluation methods

Two categories of evaluations have been identified with a variety of specific evaluation methods. For the qualitative category the main characteristic is that evaluation was of descriptive nature. In this category the evaluation outcome is dependent on the interpretation of the integration technique by the researcher(s). The quantitative category provides an evaluation based on numbers, analysis and statistics. The latter can only be achieved as the result of a successful implementation. However, the categorization of the evaluation will be dependent on the primary evaluation the research reported out on as the implementation was a means to an end. See Table 11 Evaluation methods.

2.5.3.1 Qualitative

There were two observed qualitative evaluation methods identified with key differences in the intended type of evidence to be produced. Implementation studies provides support through the actual implementation by following the proposed design and showcasing that the design is valid and will achieve set target goals. Artefact development provides support of a proposed methodology where certain activities are described and the artefacts are the means to achieve integration. For either evaluation method, the iterations and review of their work

is common. This is presented in immediate adjustments or open leads for future work (i.e. observed challenges).

2.5.3.1.1 Case Implementation

Evaluation through implementation is characterized by the presence of a use case where a (real-world) gap is presented where the solution is achieved through data integration. These cases include specific requirements to the technical implementation based on the present business environment. Implementation techniques need to adapt and overcome these challenges to achieve the benefits following data integration. Therefore, the outcome of the evaluation is linked directly to the integration technique's contribution to solve the business problem. Inability to solve the problem means it does not add business value.

Key elements present in a case implementation were observed to be of business, technical and data nature. The business context in which the problem is present in, how the problem is related to data, and surrounding constraints with regards to privacy, legal, and compliancy. It describes the relevancy of the situation and the constraints within activities need to be performed in. Technical elements can describe any technical implementation constraints such as different systems, tooling availability, any type of barrier that segregates the data. Lastly, the data elements describe the differences on a detailed data-level that need to be homogenized to identify a compatible form so interaction is possible. This can include data types, value sets, units of measurements, or context of use.

2.5.3.1.2 Artefact development

Evaluation through artefact development is characterized by the absence of a use case (as described in 2.5.3.1.1), but the presence of the development of an artefact as a result of a proposed methodology or as part of a design. This evaluation aims at solving a (generic) data integration problem through means of conceptualizing the problem and developing a solution that provides resolution. Due to the qualitative nature of this evaluation the ability to successfully develop artefact(s) to achieve the solution in itself is the evidence of its efficacy.

This type of evaluation is observed at a conceptually level where both the problem and solution is sketched in. An example, the problem is the inability to reconcile different databases and the (or a possible) solution for this is to implement a data-distributing hub that reconciles different databases. This is described in a design where potential dependencies are also outlined to achieve the solution. The evaluation occurs through the development of the elements of the design where conclusively the ability to reconcile different databases through the data hub is confirmed or not. Confirmation is always achieved however at different degrees of satisfaction. Observed constraints and improvements are necessary supporting elements to provide a meaningful evaluation.

2.5.3.2 Quantitative

There were five quantitative evaluations methods identified with key distinctions in the type of quantitative score utilized. Each method provides a different score that evaluates the data integration technique with a specific focus. This provides a straightforward comparison as long as the same score is used. However, this is only reliable if the same data set is used and possible. Similar data sets can provide guidance but essentially it cannot be guaranteed that

you are comparing apples to apples. Intra-data set comparisons are in all cases valid when comparing different integration techniques. Another constraint that applies specifically to precision and call is the necessity to know beforehand the amount of correct and incorrect results.

2.5.3.2.1 Precision and recall

Evaluation through precision and recall was the most common quantitative method of evaluation. This relies on four calculations: True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). Precision is used to measure the correctness of the results by calculating the ratio of identified correct results divided by the total number of results. A perfect precision value indicates that the produced results are reliable and with no false positives. Recall is used to measure the completeness of the results by calculating the ratio of identified correct results by the total number of correct results. A perfect recall value indicates that the produced results capture all the correct results. For integration techniques the ideal is to maximize a combination of both measures.

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

Variations were observed based on the two metrics: F-score, F2-score or Matthes Correlation Coefficient (MCC). F-score shows the harmonic mean of recall and precision in a single value where both metrics are equally weighted. F2-score extends the F-score but provides a stronger weight on recall. This score can be more meaningful, in case of an imperfect precision and/or recall, if the recalling ability is prioritized over precision. This would be scenarios where minimizing false negatives at the cost of false positives is acceptable. While F-score has no weighting and F2-score has more weight towards recall, MCC provides a score that will conclude if the results are completely random or very reliable. It achieves this through by accounting for potential bias in the data set, e.g. precision can be inflated due to the data set containing (nearly) all records to have a match during integration.

$$F\ score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

$$F2\ score = 5\ x \frac{Precision\ x\ Recall}{(4\ x\ Precision) + Recall}$$

$$MCC = \frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The metrics and scores enable a direct quantitative comparison when modifying settings or comparison of integration algorithms. With a very clear distinction what the best performer is. Additionally, the scores provide a single value to sum up the whole performance with the ability to account for weights. MCC is unique to allow for cross data-set comparison.

However, the availability of known results is a hard dependency and requires either the utilization of synthetic data or non-synthetic/real-world data. The first allows you to set the exact parameters and therefore knowing the exact number of correct results. The latter would require manual review to determine the correct results. The drawback of this evaluation is that on real-world data the number of correct results is usually an unknown factor or the need for integration would not have existed at all. Alternatively, a subset of the data can be reviewed and used for evaluation purposes. This can be very time-consuming to create and challenging to find a subset that is generalizable across the full data set. The data needs to be both unbiased in the ability to provide an accurate representation and also ensuring that not every record matches up perfect to another to bias accuracy.

2.5.3.2.2 Industry Benchmark

The Industry Benchmark evaluation method enables the comparison between different data integration solutions (from different vendors). Any data integration implementation can be evaluated as long as it incorporates certain characteristics: processing of large data volumes, various transformation steps, historical and incremental loads, reliable and accurate result requirements, multiple data sources with incompatible data formats and multiple data tables with a variety of data. These characteristics ensure that the benchmark can be technology agnostic and relevant comparisons can be achieved.

The primary performance metric for this benchmark is: number of processed records per amount of time. This single metric can be extrapolated from a partial run as it scales linearly with these specific characterized systems, i.e. a doubling of system resources will double the performance. A variation of this metric can be attained by including the price factor for comparison of the best value vendor of a data integration solution. This is achieved by measuring the required time loading rows for both historical and incremental loads. Historical loads include existing data in the sources and incremental loads are periodic updates added to the integration. Due to the frequency of incremental loads, the lowest performing load will be leading as this indicates the minimum expected performance.

$$Historical\ Performance = \frac{Historical\ Rows\ Loaded}{Historical\ Load\ Time}$$

$$Incremental\ Performance = \frac{Incremental\ Rows\ Loaded}{Max(Incremental\ Load\ Time)}$$

$$Performance\ Benchmark = Historical\ Performance\ x\ Incremental\ Performance$$

2.5.3.2.3 Scalability and resource utilization

Scalability provides evaluation by changing parameters of the data integration solution to establish if the solution holds performance when the complexity, volume of data or available resources changes. This performance could be based on correctness of the results or raw performance in integration speed. Correctness parameters are based on the number of rules and/or included attributes that determine the integration results of the data. These changes can significantly impact the computational complexity when utilizing clustering-based techniques specifically. This impact would not be as significant to decision-tree based rules.

Processing performance parameters can be more varied in both changing the characteristics of the data and available hardware resources used for processing. These changes are also more impactful to integration solutions that have are of exponential computational nature versus linear. Observed specifically with clustering-based techniques, changes in the rate of error, rate of duplicates and number of tuples provide can show if the solution is scalable or it will dramatically increase time to compute or worsened results. By tuning these parameters, an optimal balance can be identified to maximize correctness of results whilst maintaining practical processing speeds.

Resource changes shows the impact on the processing speed and can verify if there is benefit in scaling up (hardware) resources or if there are limitations to the implemented algorithms. Time to compute is key in measuring these changes. Additionally, resource utilization levels can be included to identify the efficiency of the used integration technique. With large volumes of data or volumes that exceed the capacity of a single computing machine (big data), it becomes increasingly important to ensure that resources are fully used for cost and efficient implementation purposes.

2.5.3.2.4 Comparative study

Comparisons were observed to evaluate the performance of the integration technique on different data sets or comparison of different integration techniques to the same/similar data set. This type of observed comparison does not prescribe specific score(s) to be included. Therefore, this type of evaluation is free of what scoring metrics are to be included. Comparative studies can provide validation of the proposed integration technique on its validity to a variety of data characteristics, for example different data domains, and it can also provide confirmation that there were improvements over similar/other techniques. The latter can only provide fair comparable results if the utilized data set was the same.

2.5.3.2.5 Filtering ratio

Filtering ratio provides evaluation on the ability to remove unlikely from the set of records to be processed. With large volumes of data, filtering ratio becomes a key element in achieving (continuous) integration by optimizing the efficiency of the processing. Maximizing the filtering ratio minimizes running time however the filtering needs to happen accurately to negate negative impact on the results. It is recommended to test different filtering methods and compare how which one provides the best without sacrificing accurate results. Additionally, a change in the data characteristics

2.6 Discussion

This review excluded all results from the FindUT database, a collection of research databases provided by the University of Twente. As an initial mitigation, individual databases were queried. Due to the inability to consistently apply the selection criteria and queries and the need for more consistency in the results, these databases were also excluded. As a result, only Scopus was queried and used as input for the systematic literature review. Scopus is generally considered a high-quality research source, and we expect the results to be of sufficient quality and complete to provide a good view of the field of data matching techniques. Nevertheless,

there is still a risk that some (relevant) papers might have been missed due to the exclusivity of Scopus.

A variety of data domains have been researched. In particular, People data have seen much research dedicated to it, followed by the Big Data and Product domains. Even though Big Data has been a more popularized/growing topic in the past years, there has yet to be a trend observed that this has become a focus of data-matching research. All domains mentioned above have been in equal popularity in recent years of research. However, research for 'general' master data seems to have had a gap, except for one paper in 2019. The research seems to have become less generic and more use-case driven in a particular domain. This could be explained by having historically sufficient coverage of all the conceptual, general, and high-level topics related to data integration. Besides People, Big Data and Products, the other domains seem to have few papers published. This could be due to the lack of interest/problems in these domains or more likely that this field is more use-case driven. Data integration is a means to an end, and organizations would primarily benefit from this. A purely theoretical paper would serve no added value.

There seems to be a complete gap in research focused on the supplier master data domain. This is surprising as supply chain processes, e.g. procuring, contracting, manufacturing, quality control, purchasing and payments, are dependent on a supplier master data reference to ensure each phase is transitioned smoothly, especially in organizations where the supply chain's IT landscape is distributed across different systems for each purpose. Even organizations that utilize a more streamlined system landscape can run the risk of duplicate record creation and therefore benefit from data-matching steps.

The qualitative evaluation methods are focused on implementing a solution based on the use case or proving that the proposed method leads successfully to a developed artefact of some nature. The first method seems to lack representation of the use-case's stakeholders whether or not the problem was solved. The significance of data integration could be overstated, while in the full context of real-world use cases, data integration is a means to an end. A description of required or supporting elements to successfully solve and sustain the data integration use case an organization was missed, e.g. impact of governance, change management, adopting a new integration solution, organizational buy-in, and return of investment. Ideally, research based on (real-world) use cases should attempt to provide and balance pragmatism.

The quantitative evaluation methods focused primarily on precision, recall, and variations. These metrics are fundamental in any data integration research with a known number of correct results. These metrics make it easy to see whether the results are near perfect (i.e. 100%). However, it needs to consider a meaningful threshold to fulfil a problem. Perfect precision and recall are theoretical goals. In real-world organizations, usually, 'good enough' is the goal. Usually due to diminishing added value after a certain point and/or an increase in the cost. Alternatively, some use cases could be where near-perfect is a hard requirement.

In addition to quantitative evaluation methods, there is a significant constraint on meaningful comparisons. Any calculated evaluation score must be based on the same or similar data set. This obstacle prevents a broader comparison of the performance of integration techniques. Perhaps a majority of the research is use-case driven in a particular data domain, or lack of

representable data sets for each domain could be factors that prevented more comparable research.

A broad range of integration techniques have been explored, from conceptual designs (governance and architecture) to detailed modelling and algorithmic implementations. This range of techniques can be seen as the results of a widely researched field or the complexity required to implement and sustain an integration solution fully. This shows a necessity for identifying bottlenecked elements and addressing those during integrations. Luckily, a vast pool of available literature can support each of those elements from the highest to the lowest level of detail. This does create a dependency on those users to assess each case correctly.

The integration algorithms had a clear split between machine learning and conventional record pairing techniques. Machine learning is arguably a more recent development. It provides a critical advantage that an abstraction layer is created by the machine learning model that interprets the data and provides clusters of record pairs (unsupervised) or a trained model that can pick up patterns invisible to a human (supervised). Especially useful when the data includes many attributes that need to be accounted for. In contrast to more conventional (rule-based) algorithms, e.g. string matching, this provides a critical advantage in that interpreting the results and parameter changes can be easier to understand and trace back.

Most of the literature focus on data integration from a data- or technical-driven requirement. Criteria, when records form a correct pair, are based on the determination of the integration algorithm's score. A manual review could be included that verifies the correctness of the pair beforehand or after the fact. However, there is little consideration to account for business-driven requirements. These could be in the form of business rules or logic that can impact integration steps or put specific demands on results, e.g. accounting for a set of master data records that have no room for error due to compliance, identified record pairs provide visibility but which record should be used further in the system (survivor and loser records), how to streamline the master data records within the systems and accounting for organizational/operational impact. There seems to be a focus on data integration research that mainly shows how to integrate different sources and provide centralized data access and/or analysis capability, but this means that the need for that integration solution will persist. It fixes the symptoms of a problem, but the prevention of the problem, coupled with a decreased dependency on the integration solution, is minimally unexplored. The closest thing researched was an automatic recommender on which record to have in 'the lead' based on data quality and timeliness of the latest update.

2.6.1 Implications for practitioners and for researchers

Our results provide an overview of all research data integration techniques and their challenges. This can serve as a starting point for future researchers to observe what has already been research and in what context. This can lead to further development of a particular integration technique and/or a specific data domain and how to evaluate it. Expanding on the existing work or re-using what was already applied to validate their work further. Alternatively, whatever is not covered in this review means there is a white spot and room for future work.

Future practitioners, specifically those looking for real-world applications, can use these results to identify commonalities in the research and their cases, determine which research was similar, and take the key learnings or techniques. We recommend starting in the relevant data domain or the integration techniques if the bottleneck element is known and referencing the relevant papers. This can be a pragmatic way not to get lost in theory and use what 'can work' versus the 'perfect solution'.

2.7 Conclusion

This paper systematically reviews the literature in researched data integration research between 2007 and 2020. This resulted in an overview of different data integration techniques and their associated challenges, an overview of different data domains and their characteristics, and methods of evaluations to determine performance and enable comparison. The answers to the research questions are summarised as follows:

RQ1: *What techniques have been researched for data matching and integration in MDM?* We have observed four techniques that achieve data integration: 1) Governance, 2) Architectural, 3) Integration modelling, and 4) Integration algorithms. These differ in their relative ranking in abstraction, implementation detail, and level of evaluation. Respectively, the highest level of abstraction ranks 1 to 4, the most detail of implementation ranks 4 to 1, the ability to be evaluated qualitatively exclusively ranks 1-4, and the ability to be evaluated quantitatively is exclusive to type 4. There were three data domain categories with several sub-domains: Master Data, Big Data and People Data. Master Data characteristics provide fundamental characteristics that apply to each domain and sub-domain. Big Data is distinguished by the unique hardware/compute resource requirements and People Data, including strict privacy, legal, and compliance regulations. All these techniques were evaluated via qualitative or quantitative methods. Qualitative evaluation applies for all cases, but quantitative evaluation is only viable if the number of correct results is known beforehand.

RQ2: *What are identified challenges for data matching and integration in literature*? We have observed the following main challenges. Firstly, data integration can be time-consuming and costly as it requires staffing to support various crucial data integration: result reviewing, data understanding, data quality improvement or overcoming differences in data structures or sources. Techniques that provide complete automation and/or minimise human intervention are preferred. Secondly, data quality and compatibility were constant challenges as integration is commonly required between different data sources/systems with their respective unique data models.

Additionally, 'rubbish-in is rubbish-out' applies to data quality. Correct integration cannot be achieved if data quality is already inadequate. Thirdly, preserving timeliness during changes, e.g. data volume, schema evolution, and addition of data sources, was a common challenge. This is immediately related to maximising automation and efficient data integration to reduce time and cost, especially in time-sensitive cases such as forecasting or decision-supporting system.

This review shows that there are still gaps within the research on data integration for master data management. Specific gaps that have not been explored are:

1. Application to the supplier master data domain,
2. To address the most common challenge, how can an organization implement a high-performing data integration at minimal cost due to human intervention, and
3. How can an organization improve its master data management to reduce the dependency on the data integration solution?

The first gap focuses on a missed data domain with inherently unique characteristics: the supplier master data domain. There are shared fundamentals with general master data and some elements of the People domain: contact information, location and name. However, supplier data is subject to change of the elements overlapping with People; it does not have a unique identifier globally; for legal/contracting and quality control purposes, it is essential to have a clear trace back to a single supplier master data record.

The second gap focuses on how an organization can start the journey of a data integration project with minimal costs. Master data improvements do not immediately impact an organization's bottom line, i.e. profits. It is an enabler that can be several layers away from calculating its financial impact. Therefore, it is critical to maximize efficiency. In particular, can we support organizations in minimizing time-intensive manual review on real-world data sets? This is related to also setting (realistic) definitions of a 'correct matched result' or 'good enough result'. It results in a process that takes a more pragmatic and business goal-driven approach.

The third and last gap focuses on how an organization can structurally improve its master data management post-implementation of data integration. Metaphorically, data integration is a band-aid and sustaining this keeps the organization from bleeding. However, how can an organization take the next step to improve its situation structurally to either entirely negate the necessity of the band-aid or reduce the dependency on this sustained solution? This is particularly relevant if the data integration only provides centralized visibility, e.g. for business intelligence, analytics, or traceability, instead of operational excellent where a single master data record is referenced in multiple systems, e.g. procurement, finance, quality control, or legal/contracting.

3 Research Methodology

The problem statement in the chapter above outlines a problem, solution constraints and requirements provided based on a real-world case. The goal of this thesis is to develop a solution that can address the problem with the use of data integration techniques. This is a typical design research question as we are tasked to create and evaluate an (IT) artefact to solve identified organizational problems [47]. A design research methodology will provide guidance to balance the requirements of scientific research rigour and organisational demands [48]. Therefore, we will elaborate on the selected research design and how we intend to apply the methodology.

3.1 Action Design Research

We have utilised the Action Design Research (ADR) methodology [47]. Two other design research methodologies were considered: Design Science Methodology (DSM) and Design Science Research Methodology (DSRM) [48], [49]. DSM can be characterised as an extensive framework where design research is categorised into design problems and knowledge questions with follow-up instructions on exploring them. It provides extensive and detailed tools to identify the design process, from articulating stakeholders' goals (the social context) to the exact way of formulating a design problem (the problem statement template). DSRM can be characterised as a sequential six-step methodology, with each step providing more guidelines on how to perform design research. It promotes iterative development and four starting points to initiate the design research. Arguably DSM's strengths lie in the ability to precisely prescribe the steps to perform design science from start to finish and DSRM in providing more principles to consider when performing design science in a recommended sequence of steps. Considering the formulated problem statement and stakeholders' wishes, ADR seems more inclusive to the organisation's feedback and allows for more freedom to develop the design with the stakeholders iteratively. Key advantages that we recognise over DSM and DSRM as the organisation wants more than just research.

3.2 Applying ADR

ADR defines four stages within their methodology: 1) Problem Formulation, 2) Building, Intervention and Evaluation (BIE), 3) Reflection and Learning, and 4) Formalization of Learning. From the literature review, we identified several gaps within the research that correlate with problems experienced by Philips that result in the problem formulation. Therefore, the literature review will provide the theoretical basis and input to the design of the artefact(s) for the BIE stage, which correlates with the IT-dominant BIE start. This is where the start of the artefact development/design is based on the researcher's input, and we will initially propose an artefact based on the literature as there is no existing artefact to extend.

Stakeholder interaction is a key factor that needs to be included in the development of the solution. For the reasons stated in Sections 1.1 and 3.1, the aim is to maintain scientific rigour and achieve real-world added value for our stakeholders. Therefore, we need to emphasize and clarify who the stakeholders are, their expected input/output, and how we will interact with them. We describe this in the ADR methodology as practitioners and end-users. They will be interacted with to develop the alpha and beta versions of the solution, respectively.

Philips utilizes a joint collaborative governance on data, IT systems and processes to enable business capabilities such as the Purchase to Pay process, see Figure 8 Stakeholder governance . Three generic roles can be described in this governance structure but the exact implementation might differ within reality depending on the domain.

- Business Process Experts (BPE): stakeholders positioned to focus on the data and the business process. They focus on designing and implementing a correct business process with the requirements of what data is needed to enable this. Ideally, the process design activity can be characterized as system-agnostic, meaning that the process is not dependent on the tool, and the tool is configured as such to enable the process. In reality, they will have to account for the existing IT landscape to some extent, as simply replacing a system is costly and are long-term projects.
- Business Information Experts (BIE): stakeholders positioned to focus on data and the IT systems. They focus on ensuring that the IT system(s) interfaces with each other correctly, ensuring that data is available and of good quality to enable the business process. They play an important role together with the BPEs to harmoniously design a holistic package that enables business capabilities as they focus on ensuring the requirements from a data/IT perspective are accounted for.
- Subject Matter Experts (SME): stakeholders positioned to focus on the IT system and the business process. They focus on the delivery of the capability of the business capability, i.e. they are the operational support for the users within a capability, such as buyers in the Purchase to Pay process. Due to their close proximity to the users, they have a unique operational perspective that can identify if the outcome of the complete package (data, system, process) is adequate.

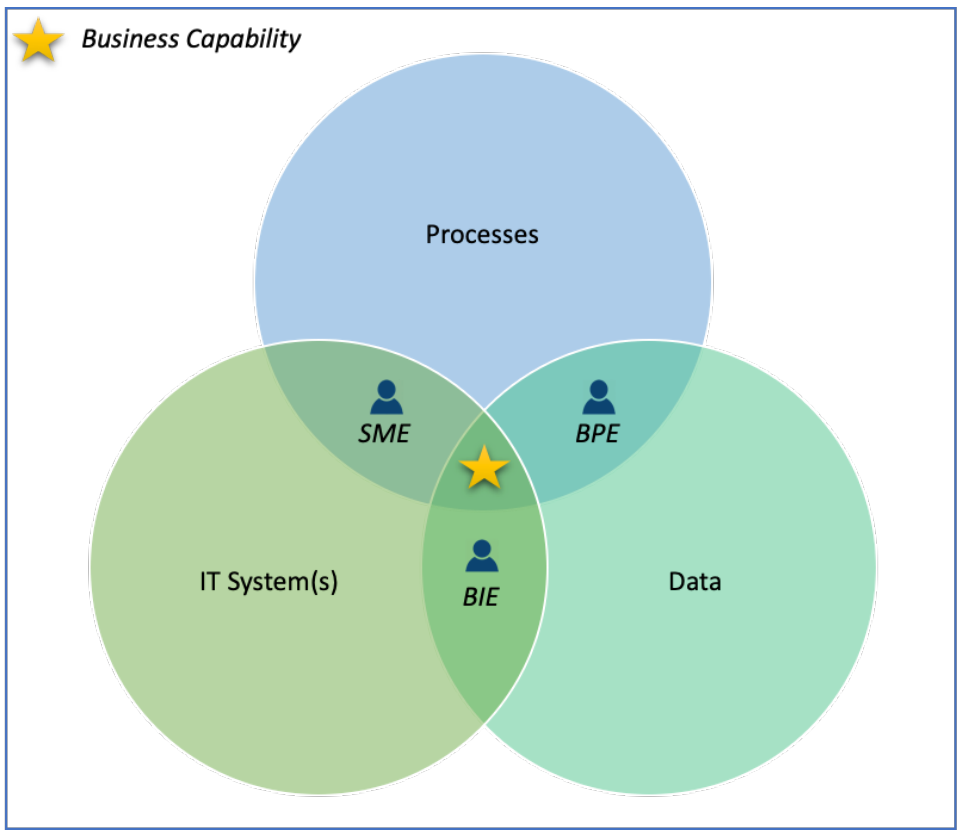

Figure 8 Stakeholder governance structure

For the development of the alpha version of the artefact, we will iteratively develop the artefact together with the BPEs and BIEs. They will engage with us to ensure that the solution direction complies with the existing process and that the solution can be applied within the existing (available) data and systems. The alpha version will prove that the solution artefact in concept will be able to resolve the observed challenges. To develop the beta version of the artefact, the SME will provide their input and evaluation to ensure that the expected outcome of the solution artefact is valid. Their knowledge ensures that the intended design will work to the level of day-to-day operational activities. We do not include the actual end-users of the Purchase to Pay process, i.e. the buyers, because the positioning of the SMEs as a supporting role means they will have the same core knowledge as a buyer and additionally have an awareness of (common) issues/risks. Therefore, they can provide a far more in-depth and meaningful solution validation. An overview of this interaction can be found in Figure 9 Overview of the BIE-phase and expected outcome.

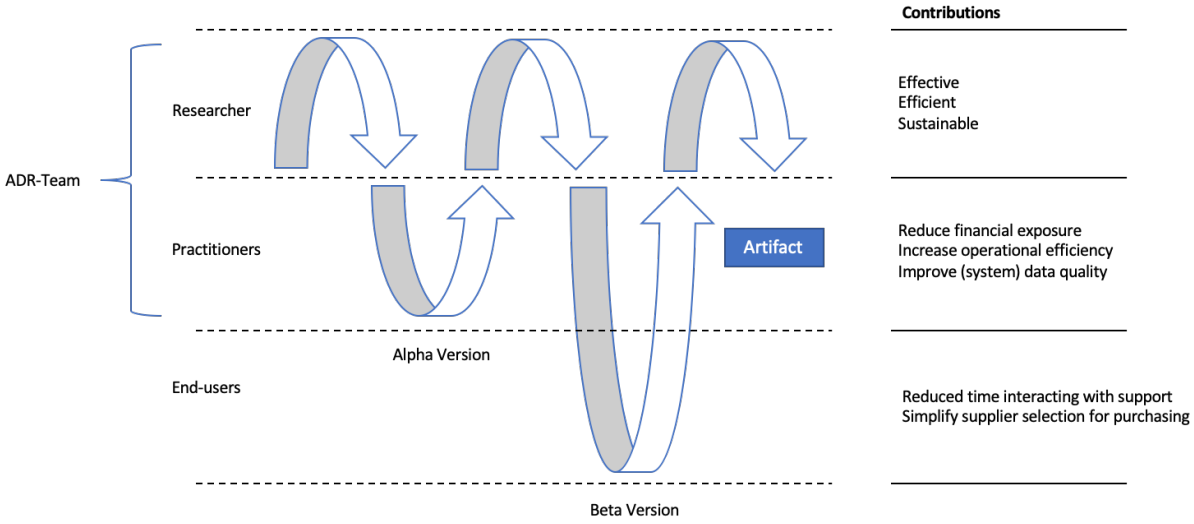

Figure 9 Overview of the BIE-phase and expected outcome

3.3 Expected contributions

Following the methodology of ADR, the contributions can be categorized into three areas depending on the beneficiary: 1) design principles, 2) practice-specific contributions, and 3) user utility. The design principles are generalized extractions from the final solution artefact that allows researchers to apply common concepts to their problems, effectively casting this instance of a problem into a class of problems where the design principles can address this class of problems and add further to the field of research [47]. The practice-specific contributions and user utility focus on this solution artefact's benefit to the organization and their user community.

The exact formalization of the contributions will occur after the development of the artefact and reside in step 4, see. However, based on the initial solution requirements provided by the organization in **Error! Reference source not found.**, we can already ideate some guiding p rinciples and expected outcomes that will support the BIE-stage two-fold: 1) guiding principles for the solution design and 2) evaluation criteria for the resulting design. The first supports the researcher with the initial theory-based artefact proposal in step 1, and the second can

support the (iterative) evaluation between steps 2 and 3 by providing the stakeholders with an overview of the expected added value.

3.4 Summary of the Research Methodology

Figure 10 Implementation of the Action Design Research Methodology shows an overview of the implementation of the methodology utilized in this thesis. The literature review in Chapter 2 provide a theoretical starting point for the artefact design. Further development and validation will be performed with three key stakeholders: 1) Business Process Experts, 2) Business Information Experts, and 3) Subject Matter Experts. Respectively, they provide input for process, data and actual use validation.

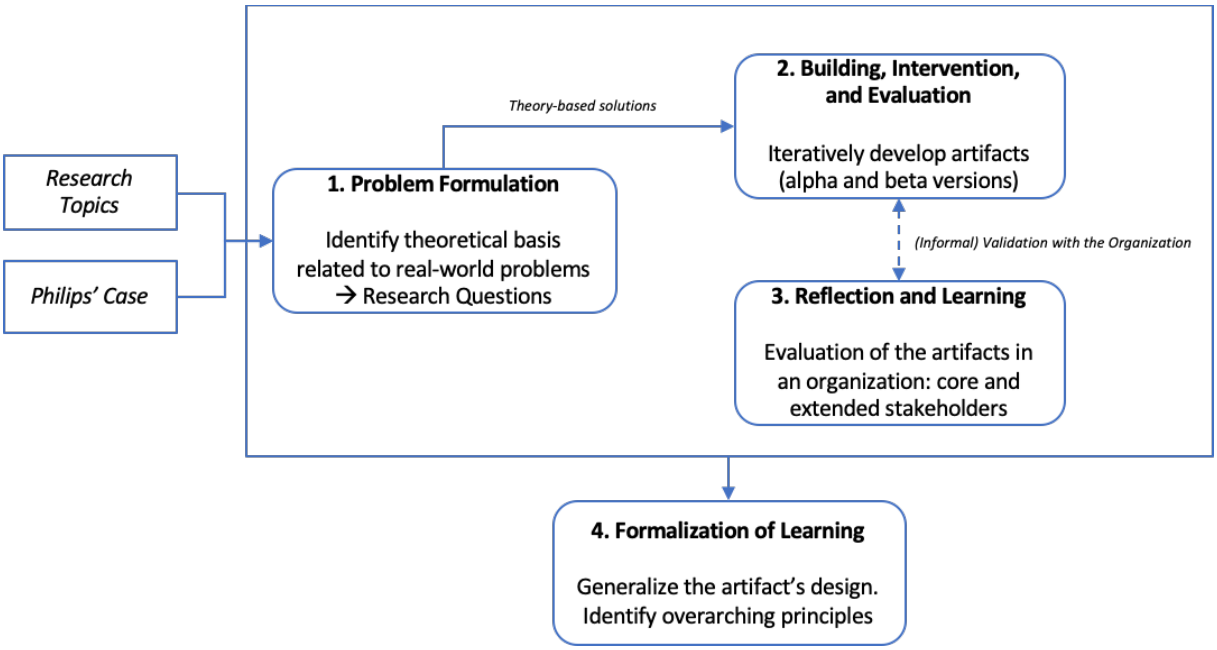

Figure 10 Implementation of the Action Design Research Methodology

4 Results

This chapter combines the results for both sub-research question 1 and 2. We found several algorithms that have been researched and applied to different data domains in the literature review. Unfortunately, there was no specific research applied to the SMD domain. Therefore, we will compare integration algorithms and measure their performances. We will first select relevant integration algorithms based on the research domains that show similarities with the SMD domain.

4.1 Supplier Master Data

Within the literature review in 0, we have provided an overview of the researched data domains and observed characterizations of each domain. Based on the dataset we sampled from the organization, we see similarities with the general Master Data and People Data domain, specifically within the Customer sub-domain, as the generic Master Data domain provides the main characterizations of being the fundamental core set of data that connects core data utilized in operational activities, linking systems and allowing for centralized data management. SMD fulfils the same characteristics with the only addition of incorporating supplier-specific data. This data contains information about organizational entities in the market for providing a service and/or goods as a transaction with purchasing customers.

There are also similarities with the Customer Data sub-domain of the People Data domain. Customer data generally include attributes such as name, address, contact information (e-mail or number) and banking information/reference to a governmental service. These attributes are a commonality with the dataset we will utilise in this research, see Section 4.1.1, and with the general understanding of supplier data.

However, there are some key distinctions with Customers in name and legal compliance attributes. Firstly, the name suppliers can be considerably more creative, redundant, simple and complex simultaneously. The names of entities can be almost freely chosen, and based on this, we sometimes see references to the name of an individual also freely selected. Including a legal entity reference such as “NV” or “BV” for the Netherlands is common practice. These are key differences and need to be accounted for in integration activities as these serve limited purposes to the uniqueness of an organization. There is an exception where an organization is sufficiently large and has a parent and children legal entity relationship.

Additionally, supplier names can be very simple or complex. Organizations such as “3M” or “Muenchener Rueckversicherungs Gesellschaft in Muenchen AG” have existed. Secondly, there are different regulations we need to abide by to handle customer data that refers to individuals. This would have affected how you process data, e.g. getting consent to use your data for a specific purpose. Fortunately, Personal Identifiable Information regulations do not apply to businesses, and we do not need to take extra measures as long as we do not include individuals’ contact information (e.g., an employee's name).

Based on the characteristics of the described domains, either in intended use or overlap in attributes, the SMD domain has shown similarities. Therefore, the researched algorithms for the Master Data and People Data domain are relevant for integration within the SMD domain.

4.1.1 RIDDLE Restaurant Dataset

For the evaluation of integration algorithms, we require a dataset of Supplier Master Data. Developing a proprietary dataset will be time-consuming, and the organization does not have a dataset on hand that we can utilize that includes labels to indicate which records are duplicated from each other. One of the discussed observations within the literature review was also the lacking comparative power between researches due to differences in the utilized data, see Section 2.6. However, there have been attempts to resolve this lack of comparison ability. The University of Texas has attempted to create a centralized repository of datasets for duplicate detection, record linkage and identity uncertainty, i.e. data integration activities. This dataset has been used at least once, which shows that it is viable, and we will have a comparison once we have implemented the algorithms ourselves [50].

The RIDDLE dataset contains restaurant records that were obtained from the Zagat and Fodor restaurant guides. These include information about the name, street, house number, city, telephone number and restaurant category. The street, house number and city have been grouped as address and city attributes. This is related due to how the information was provided. A sample of the data can be found in Table 2 Sample of the RIDDLE dataset. There are in total 864 records, where 534 originated from Fodor and the remaining 330 from Zagat. It includes a total of 112 duplicate pairs that relate to an equal number of records per source. These 112 duplicate pairs will be the leading information to assess performance in the integration algorithms.

Name	Address	City	Telephone	Category
arnie morton's of chicago	435 s. la cienega blv.	los angeles	310/246-1501	american
arnie morton's of chicago	435 s. la cienega blvd.	los angeles	310-246-1501	steakhouses
art's delicatessen	12224 ventura blvd.	studio city	818/762-1221	american
art's deli	12224 ventura blvd.	studio city	818-762-1221	delis
hotel bel-air	701 stone canyon rd.	bel air	310/472-1211	californian
bel-air hotel	701 stone canyon rd.	bel air	310-472-1211	californian

Table 2 Sample of the RIDDLE dataset

4.2 Integration Algorithms

An overview of integration algorithms researched per data domain can be found in Table 12 Overview of integration algorithms per domain. The integration algorithms from the Master Data and People Domain will be in-scope for further comparisons. This chapter will elaborate on the different types of categories and their main characteristics, pros and cons to achieve data integration.

4.2.1 Edit distance-based

Edit distance-based algorithms are fundamentally based on the string comparison category of algorithms. String comparison algorithms, at its core, compare strings to each other and evaluate the likeliness of how similar they are. Edit distance achieves this by performing character-by-character comparisons through addition, deletion and substitution (a character change) [51]. The fewer change operations required, the more similar the strings are to each other. The exact scoring value depends on the algorithm's implementation, such as Levenshtein or Hamming distance. These can include additional conditions such as only equal string lengths in the case of Hamming distance, where Levenshtein allows for all lengths [52], [53].

The main advantages of edit distance-based algorithms are flexibility, (short) string comparison strength, and ease of implementation [54]. Firstly, these algorithms can be applied to any data if represented as strings. It can be applied to all the types of attributes that we observe in the RIDDLE dataset. There are no limitations on the minimum or maximum string length. Therefore it can be applied to different datasets without issues.

Secondly, due to its character by character-based comparisons, it can handle typos and short-string comparisons well. Where two strings differ due to the accidental inclusion of a character or misspelling, which is reasonable for any type of Master Data due to human administration, it can account for the (minor) difference and conclude they are still similar. For example, this would apply to the strings “Philips” and “Phillips”, where the difference is only one operation.

Lastly, edit distance-based algorithms are reasonably easy to implement and understand. There are numerous open-source implementations of the algorithms available where one only would require to write the logic to apply the comparison. Additionally, its simple comparison concept makes it easy to evaluate and understand why the algorithm has outputted a particular score compared to machine learning algorithms, where there can be a lack of understanding and transparency on how the score has been calculated (a black box).

The main disadvantages are sensitivity to string length, large computation and lack of classifications [54]. Firstly, we previously mentioned that there is a strength in (short) string comparisons, but this is a double-edged sword. Due to its character-based comparison, there can be situations where a single letter difference is a distinguisher. For example, names that include Dutch legal entity references “NV” and “BV” are distinct legal entities and are, by definition, not the same supplier (disregarding parent-children relationships).

Additional constraints must be included, such as excluding these legal entities or making the number of operations relative to the string size that determines the score. Secondly, many comparisons need to be run, i.e. all the characters. This must be accounted for, especially when the strings become larger, as time is a valuable resource for organizations. In scenarios where time is scarce or timeliness is critical, these algorithms require additional care in implementation, e.g., more potent hardware for computing or the introduction of blocking.

Lastly, one of the key disadvantages of edit-distance and general string comparisons is the lack of classification it provides. The algorithms only provide a score on the similarity between strings. At the upper and lower ends, it is clear that there is either a match or no match. However, in between, a wide range of scores requires setting a threshold to determine whether the records match or provide more classifications such as maybes. Because these are estimations, you will leave the door open for false positives and true negatives. Minimizing these two types of results can be time-consuming and imperfect due to the balance one usually needs to strike.

4.2.2 Token-based

Token-based algorithms can also be described as general string comparisons. The main difference with edit-based is that the evaluation happens on the token level. Through tokenization, where a string is split into tokens, comparisons occur between the tokens of the two strings [55]. For example, the strings “hello” and “world” can be broken up as “hel”, “lo”, “wor” and “ld”. The exact tokenization can be based on the implementation and/or

configurations, i.e. token lengths and overlap. During token-based comparisons, the tokens will be used to see if they occur in the to-be-compared string. In the example, none of the tokens exist in the opposite strings. A popular algorithm of this type is the Jaccard index. It considers the strings as sets where the union determines the similarity.

The advantages are the weighting of tokens, contextual understanding and computational efficiency [55]. Firstly, the tokenization process allows for identifying more common tokens in the to-be-compared strings or full dataset. This allows for putting more importance on unique tokens and assigning lower-scoring importance to common tokens. This is especially valuable in longer strings such as addresses. The importance of the token “the” in everyday language is less important to determine the meaning of the text.

Secondly, contextual understanding is an important factor in determining similarity. For example, in the previous example of “NV” and “BV”, the strings can be tokenized with length 2. The comparison would conclude that there is no similarity where a naïve edit-distance algorithm might score it 50% similarity due to requiring only 1-character operation of the total 2.

Lastly, depending on the tokenization length, it can significantly reduce comparisons that are required to compute. Improving the required time to run the compute and saving time/costs. An important factor for organizations when integrating data.

The disadvantages are length of token selection, short string sensitivity and lack of classification [55]. Tokenization requires the declaration of the token size. The token size has a direct impact on the similarity scores. Therefore, it is important to perform experiments to achieve the best size that maximizes true positives and true negatives. This can be a time-consuming activity, and preferably it is already known what the correct matches are to calculate performance easily.

Secondly, we previously mentioned that the ability to distinguish short strings when they are truly distinct from each other is a pro. However, in the case of two short strings that are true matches, this type of algorithm concludes they are non-matches. This could be a specific scenario where the short string has a typo. This could be the case of a naïve implementation and the strings “Apple” and “Aple”. Additional logic would be required to mitigate this, such as overlapping tokenization.

Lastly, similar to the edit-distance-based algorithms, no classification is outputted. Identifying a threshold to determine matches and non-matches requires the same activities as before.

4.2.3 Supervised Machine Learning (Classification)

Machine Learning (ML) has become a popular field of research that utilizes trained models to provide answers to classification problems. Data integration inherently requires matching records to each other, i.e. are the records a match or not? This type of question can be resolved using ML [56], [57]. Supervised ML requires explicitly a training and test set of data. This includes records with labels indicating whether or not they are matching (to a different record) or non-matching. The training set serves as a ‘cheat sheet’ for the model to identify (hidden) patterns resulting from the provided classification label. The resulting trained model will then apply their learnings on the test set to validate its performance. These patterns are

presented in the way of features. Features can be a single or collection of attributes with specific values common in a particular classification. A model's exact training depends on the utilized algorithm, such as Naïve Bayes or Support Vector Machines (SVM). Where the first looks at features independently, and SVM can look into interactions between them.

The advantages are (hidden) feature recognition, generalizability, continuous improvement and classifications [58]. Firstly, feature recognition is a significant benefit of machine learning. It can pick up on (hidden) patterns where a human takes considerable time to deduce them or not even pick up on them. This commonly takes the form of rule-based classification where conditions need to be met. However, when the number of attributes or data size becomes more extensive, a straightforward rule-based approach might be limited in results and increasingly challenging to define.

Secondly, once a model is trained, it can be applied to new data as long as the training data represents the larger set. Therefore, a trained model can be effectively deployed without seeing the full extent of the data and not having to train on all the data.

Thirdly, a model can be re-trained or further trained to improve performance. Suppose there are indications that the model is not performing classification as intended. In that case, the same process can be followed with new training and test data sets without significantly changing the initial process. Continuous training of an existing model is also an existing field of research called reinforcement learning. This is a more advanced form of (continuously) training a model where the output is evaluated positively or negatively. It serves again as input for the model to adjust its parameters. Lastly, the output of a supervised ML model is a classification. This negates the necessity of creating your classifier, as with the previously mentioned string-based comparison algorithms.

The disadvantages are the creation of labelled data, feature selection, sensitivity to bias and interpretability[58]. Firstly, there is a hard dependency on having training and test data available that is representable for the complete set. Organizations usually do not have this available or in a small size. This can be a time-consuming (manual) and costly activity to create this from scratch. Either an organization already has this data which would mean they already have a way of identifying matches and are in the progress of doing this, i.e. they have a way of integrating data, or they have spent much time to gather this information, i.e. this can be a very costly activity. Therefore, gathering this information takes time and effort.

Secondly, selecting the right features that are important for the classification can be difficult as it relies partially on the known insights of the experts, but this information is not necessarily complete. Therefore, the activity of feature engineering can be a process of trial and error. Thirdly, combined with the previously mentioned disadvantages, if the selected data has been trained and feature selected, it can still result in weak generalizability on the larger set. This is also known as overfitting, where the training data is limited, so all the learnings the model does will over-exaggerate the exact patterns of the provided input.

Lastly, ML models are inherently difficult to interpret because they can process large complex data sets and derive (hidden) patterns from them. This can create a black-box situation where step-by-step tracing back how a model concluded is impossible in contrast with rule-based approaches or more intuitive string comparisons.

4.2.4 Unsupervised Machine Learning (Clustering)

ML models can also be trained in an unsupervised manner. In contrast to supervised ML, unsupervised ML does not require any training or test data set as it does not create a model with learnings on patterns. It creates a resulting clustering of data points where closer groupings indicate stronger similarity, i.e. a dense cluster. Matching records of a distinct supplier would be represented as a cluster in this way [50], [59], [60]. It achieves this similarly by identifying patterns within the data through its attributes resulting in features. However, the main difference is that unsupervised ML does not attempt to find a pattern within a group of prescribed clusters, i.e. the labels. The way of clustering depends on the utilized algorithms. For example, k-means and hierarchical clustering are distinct from each other as the first achieves this by pre-defining the number of clusters it needs to create, and the latter follows either a top or bottom-down approach where clusters can be merged or split.

The advantages are flexibility, (hidden) feature recognition, and classifications through clustering [58]. Firstly, unsupervised ML does not require the existence of specific attributes or labelling. It can generally interpret data of different complexities, e.g., more data attributes, without the need for labels. Depending on the algorithm, it does require specific pre-processing steps. Based on this, unsupervised ML would be feasible on different data domains or underlying sub-domains.

Secondly, similar to supervised ML, recognising (hidden) patterns is a significant benefit of ML techniques. It provides the same advantages as mentioned before and without the necessity of labels. Therefore, it is entirely independent of the need for existing insights from experts to guide the clustering where there is a risk of faulty insights too.

Lastly, the resulting clusters translate to the classification of matches as distinct entities. This is a general advantage of ML over the previously mentioned edit-distance and token-based algorithms.

The disadvantages are lack of validation truth, sensitivity to bias, and interpretability [58]. Firstly, the lack of dependency on labelled data implies no validated truth to compare the clustering results. Therefore, there will be a risk of clustering that includes too many or too few records from the actual truth. Creating this type of truth brings back the need for activities to label some parts of the data manually. The larger the validation, i.e. the size of labelled data, the more time and costs.

Secondly, bias is also a disadvantage to this ML approach. Unlike supervised ML, this type of bias is based on including data that provide noise, preventing the identification of relevant features. For example, if attributes are included that default to a specific value, these can be interpreted as an underlying pattern. Therefore, it is important to scrutinize the data and disregard not meaningful attributes.

Lastly, a recurring disadvantage with ML algorithms is the need for more transparency and traceability of how results are produced. This is the case with unsupervised ML, where the algorithm works to identify (hidden) patterns that can introduce a black box of decision-making. This makes it unsuitable for applications where a clear auditing trail and detailed explanation are required on how a particular result is achieved.

4.2.5 Other

One other integration algorithm was observed: File Integration Strategy. This prescribed general steps on how to identify and log and merge files. By identifying the requirements and legal compliance dependencies, a guideline can be developed for handling the files that need to be integrated. This is followed by extensive logging of what particular file has been merged and mappings of variables that represent various attributes of the files, i.e. similar file names or type of file format. Lastly, the merge activity occurs in the form of files being merged within the same folder. This was applied in medical research where data integration is the grouping of relevant medical files. Due to the nature of integrated files and folders, we exclude this from the comparison of algorithms as this does not apply to the integration of supplier data records.

4.3 Algorithms comparison

We compare integration algorithms on the RIDDLE dataset and their performance. For the comparison, we have excluded the supervised and unsupervised ML algorithms as the RIDDLE dataset has only a limited number of attributes that would result in a strong overfitting of the data. The researchers behind the dataset reported that including telephone attributes would already result in a leading matching indicator. However, this field in the Philips dataset is not equally available or maintained.

4.3.1 Data preparation

The RIDDLE dataset was extracted from an Attribute Relation File Format (ARFF) and a plain text file (TXT). The records were provided as plain text in both, where the first format provided a structure to extract the different attributes, e.g. name, address and city. However, this file did not include a reference to its source, which was required to treat the data as two different sets and allow the comparison algorithm to evaluate between them. The TXT format referenced the source but required some pre-processing before we could link the source. The two file formats did not match perfectly on the same expected record. Therefore, a manual review was required to check for the (slight) differences and assign the correct source.

The resulting data structure can be found in Table 2 Sample of the RIDDLE dataset. This shows that address is the concatenation of street and house number. Based on expected data formats from the organisation, we have simplified the attributes to simplify the number of comparisons. This resulted in the address field that combines street, house number and city. This was done for future-proofing as the setup does not need to account for the different structures of supplier data from the organisation's systems. The inclusions of fields such as region and state can all be added to the combined address field.

The researchers behind the RIDDLE dataset already recommended excluding the telephone attribute from any comparison research. Upon further review of the attributes, the name and address fields were deemed the strongest indicators of uniqueness. Therefore, we have dropped the category attribute from the comparison as the previously mentioned attributes would be in all casing stronger indicators. For example, a mismatch in name and address while the category matches still results in a non-match.

4.3.2 Setup

The implementation was entirely built in Python code. We have leveraged an existing public library called TextDistance. This library is provided under MIT licensing conditions that allow us to utilize it in full without permission requirements. TextDistance is a library that provides a large number of distances comparing algorithms. It incorporates all the integration algorithms identified in Table 12 Overview of integration algorithms per domain. Each integration algorithm was used in its pure Python implementation to provide an equal playing field for the time to compute comparison. For all algorithms, the normalization variation was utilized in calculating the scores. This allowed the scores to all range from 0 to 1, where 1 would be a perfect similarity match and 0 a complete mismatch.

The dataset was prepared as two Pandas Dataframes (DF). Iteration logic was wrapped around the DFs to perform row by row comparison between the two. This resulted in a comparison of 330 (Zagat) * 534 (Fodor) = 176.220 records for a single algorithm. The total number of comparisons would result in 1.409.720 resulting records that would translate to an equal number of outputting rows. Due to the post-processing and analysis occurring in Excel, the output had to be limited due to restrictions on the maximum number of rows Excel could handle. Therefore, all record pair comparisons with a score below 0,1 were excluded. These records are on the lowest end of the similarity score range and are expected not to impact the resulting matches' scores.

All record pairs are calculated between the two DFs to calculate the Cartesian product. We require all record pair similarity scores as the string comparison algorithms do not provide a classifier. Therefore, our own classifier has to be built and pre-calculating all possibilities enables the analysis step to identify the best threshold for performance comparisons.

The time to compute comparison depends on the implementation and the hardware where the execution took place. As the implementation logic remained equal between the different algorithms and the hardware, too, the times would be comparable within this setting. The execution took place on a system with Windows 11, AMD Ryzen 5 PRO 6650 U and 16 GB of memory. The implementation did not include any multi-threading workload distribution, blocking optimization or any other dedicated optimization.

Configuration	Specification
Operating System	Windows 11
CPU	AMD Ryzen 5 PRO 6650U
RAM	16 GB
TextDistance	4.5.0
Python	3.10.9

Table 3 Setup configuration details

4.3.3 Analysis

Initially, a classifier has to be built to determine whether or not the record pair scores for name and address resulted in a match. More classification is possible however, this is mainly done for a manual review follow-up which is out of scope of the analysis. The setting of a

match threshold can result anywhere between the 0 and 1 scores. However, to provide a more statistical approach that is objectively applicable in any future integration setting, we will consider the scores' minimum, maximum and mean to define our threshold. The ideal outcome would be to maximize true positives and minimize false positives. This would be achieved by testing all threshold scores and comparing the aforementioned numbers. However, in our case, the matching record pairs are known, and therefore we will leverage these scores to determine a threshold to be applied to the entire set.

The outcome of the classifier will be a match or non-match result between all record pairs. We will perform a quantitative analysis that will use precision and recall as the key metrics. The F-score will be used to directly compare the algorithms where the highest score will be the best-performing algorithm. The additional scores will also be computed to provide additional insight when different weights of importance are applied, see Section 2.5.3.2.1 Precision and recall.

In addition to performance through precision and recall, we will also include a time to compute metric. This is the time required for the algorithm to start and compare the two data sets. This can provide additional information to an organization implementing an integration solution for activities such as planning and resource allocations. In generally all cases, it is expected that differences in time to compute are negligible as long as performance is high. However, the exception is in case the time to compute several factors is higher that it becomes impractical for real-world use.

4.3.4 Results

The initial classifier needs to be built to configure a threshold to identify the matches and non-matches further. We will cover the configuration first and use that to generate the results of the comparison.

4.3.4.1 Classifier Configuration

The initial results for building the classifier are based on the records that are known duplicates of each other. This resulted in scores for the name and address per record pair and algorithm. An overview of the standard deviation, minimum, maximum, and mean can be found in Table 5 Statistics overview of the known matching record pairs and visualized in Figure 11 Distribution of scores per algorithm. The minimum value of all the integration algorithms is considerably low. They range from the lowest name scores of 0,1 with Hamming Distance and 0,6 with Jaro-Winkler, and lowest address scores of 0,1 with Hamming and 0,5 with Jaro-Winkler. It indicates that even though these are confirmed matching record pairs, the algorithms could not provide a high similarity score. Based on sampling, we have identified that the number of false positives for these minimum scores would be very high, see Table 4. Therefore, a threshold based on the statistical minimum is discouraged due to poor performance.

Name_Zagat	Address_Zagat	ID_Fodor	Name_Fodor	Address_Fodor	Score_Name	Score_Address
art's deli	12224 ventura blvd.studio city	3	art's delicatessen	12224 ventura blvd.studio city	0,9	1,0
bel-air hotel	701 stone canyon rd.bel air	5	hotel bel-air	701 stone canyon rd.bel air	0,6	1,0
fenix at the argyle	8358 sunset blvd.w. hollywood	15	fenix	8358 sunset blvd. westhollywood	0,9	1,0
le chardonnay (los angeles)	8284 melrose ave.los angeles	25	le chardonnay	8284 melrose ave.los angeles	0,9	1,0

Table 4 Examples of false negative results with the mean threshold (JaroWinkler)

A maximum value can also be considered for setting a threshold. The observed maximum value for all algorithms was 1,0. This would intuitively result in only those matches with the highest confidence due to the maximum possible similarity score. This threshold minimizes false positives but also limits the number of matching record pairs that are (very) close in similarity, i.e. increasing false negatives. Alternatively, the mean can also be considered as it would provide a less extreme position on the required similarity score. Both will be used as thresholds in the classifier to identify which one provides the best precision and recall.

	Attribute = Name						
Statistic	**Cosine**	**Damerau**	**Hamming**	**Jaccard**	**JaroWinkler**	**Levenshtein**	**SmithWaterman**
std	0,1	0,2	0,2	0,2	0,1	0,2	0,2
min	0,4	0,2	0,1	0,2	0,6	0,2	0,2
mean	0,9	0,9	0,9	0,9	1,0	0,9	0,9
max	1,0	1,0	1,0	1,0	1,0	1,0	1,0
count	111	110	100	111	111	110	97
75%	1,0	1,0	1,0	1,0	1,0	1,0	1,0
50%	1,0	1,0	1,0	1,0	1,0	1,0	1,0
25%	1,0	1,0	1,0	1,0	1,0	1,0	1,0

	Attribute = Address						
Statistic	**Cosine**	**Damerau**	**Hamming**	**Jaccard**	**JaroWinkler**	**Levenshtein**	**SmithWaterman**
std	0,1	0,2	0,3	0,2	0,1	0,2	0,3
min	0,5	0,1	0,1	0,4	0,5	0,1	0,1
mean	0,9	0,8	0,7	0,8	0,9	0,8	0,7
max	1,0	1,0	1,0	1,0	1,0	1,0	1,0
count	111	110	100	111	111	110	97
75%	1,0	1,0	1,0	1,0	1,0	1,0	1,0
50%	0,9	0,8	0,8	0,8	1,0	0,8	0,8
25%	0,8	0,6	0,5	0,6	0,9	0,6	0,6

Table 5 Statistics overview of the known matching record pairs

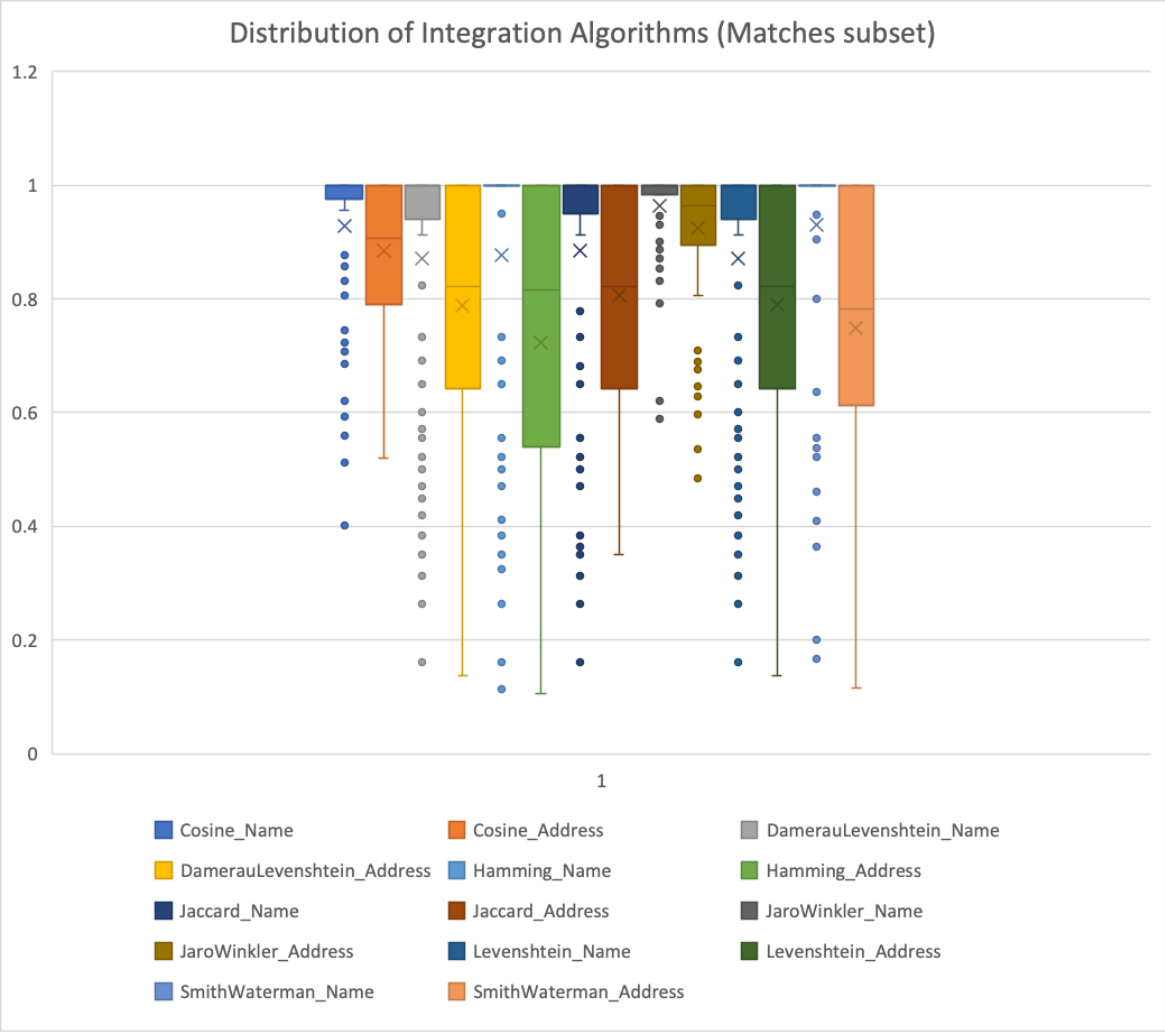

Figure 11 Distribution of scores per algorithm

4.3.4.2 Quantitative comparison

The results of the performances can be found in Table 6 Overview of performances of the algorithms. Based on the two analyzed thresholds, it is clear that the recalling power of the mean threshold was for all the algorithms better than a max threshold classifier. This is without the sacrifice of precision and shows that the max threshold classifier would miss the matching record pairs with very high similarity scores. This threshold setting is a maximum score, and it shows limitations in its ability to recall the true positive pairs that are similar but not identical. This is reflected in the variation of the recall scores in the mean classifier compared to the max threshold classifier, where only two unique recall values are measured. Forcing a max threshold seems to minimize the differences between the algorithms, whilst they uniquely can have strengths to identify matches. Going for this approach makes the algorithm selection minimal in terms of performance.

The mean threshold classifier improves recall across the board whilst maintaining perfect precision. Specifically, the JaroWinkler, Cosine and Jaccard algorithms saw the biggest performance increases for all the scores with an average increase of 0,1. These seem to have the strongest positive relationship with lowering the threshold. The Damerau Levensthein and Levenshtein algorithms did not increase their recalling performance with the same power, but

they did peak in the MCC score. This indicates that when considering all possible four outcomes, e.g. true positives, false negatives, it provides the best-balanced performance overall, whilst JaroWinkler has the best recall performance.

Mean threshold

Algorithm	False Negative	True Negative	True Positive	False Positive	Grand Total	Precision	Recall	F-score	F2-score	MCC
Cosine	78	81	34	0	181	1,0	0,3	0,5	0,4	0,4
DamerauLevenshtein	81	224	31	0	323	1,0	0,3	0,4	0,3	0,5
Hamming	85	10	27	0	97	1,0	0,2	0,4	0,3	0,2
Jaccard	78	60	34	0	159	1,0	0,3	0,5	0,4	0,4
JaroWinkler	72	55	40	0	154	1,0	0,4	0,5	0,4	0,4
Levenshtein	81	223	31	0	322	1,0	0,3	0,4	0,3	0,5
SmithWaterman	81	32	31	0	123	1,0	0,3	0,4	0,3	0,3

Max threshold

Algorithm	False Negative	True Negative	True Positive	False Positive	Grand Total	Precision	Recall	F-score	F2-score	MCC
Cosine	87	81	25	0	193	1,0	0,2	0,4	0,3	0,3
DamerauLevenshtein	88	224	24	0	336	1,0	0,2	0,4	0,3	0,4
Hamming	88	10	24	0	122	1,0	0,2	0,4	0,3	0,1
Jaccard	87	60	25	0	172	1,0	0,2	0,4	0,3	0,3
JaroWinkler	88	55	24	0	167	1,0	0,2	0,4	0,3	0,3
Levenshtein	88	223	24	0	335	1,0	0,2	0,4	0,3	0,4
SmithWaterman	87	32	25	0	144	1,0	0,2	0,4	0,3	0,2

Table 6 Overview of performances of the algorithms

We have also gathered the time to compute for the RIDDLE dataset and estimated the necessary computing time for varying data sizes, see Table 7 and Figure 12. Figure 12 excludes the compute time for the SmithWaterman algorithm because it is significantly higher in compute time than the others which makes it unfavourable for use. Overall, the time to compute on the RIDDLE data set was in the range of seconds for 176.220 comparisons. We used this number to estimate how much time would be required if the comparisons increased through larger datasets. This gives an indication not of the exact projected time to compute required but an estimation of the range of time it requires.

Algorithm	RIDDLE (n=330, m=534)	n=10000	n=20000	n=30000	n=40000	n=50000	n=60000	n=70000	n=80000	n=90000	n=100000
Levenshtein	22	3	14	31	55	86	124	169	221	280	346
JaroWinkler	20	3	13	28	50	79	113	154	202	255	315
Jaccard	38	6	24	54	96	150	215	293	383	484	598
Hamming	24	4	15	34	60	94	135	184	240	304	375
Cosine	32	5	20	45	80	125	181	246	321	406	502
DamerauLevenshtein	21	3	13	30	54	84	121	165	215	273	337
SmithWaterman	491	77	310	697	1239	1936	2788	3794	4956	6272	7744
	(seconds)	(hours)									

Table 7 RIDDLE and expected compute times

We observed that the SmithWaterman algorithm takes 13 times longer than the second-slowest algorithm (Jaccard). Due to this, we have excluded this algorithm as, from purely a time perspective, it is not practical for use. The other algorithms differ from each other within seconds with the RIDDLE dataset. However, the differences become significant once we projected time to compute with hypothetical increases in data size and compute, n = size of the data where the compute is n times n.

With a size between 20.000-30.000, the time requirement is around 24 hours before full completion, and 60.000-70.000 requires around a week. This is an important threshold for the maximum time to compute can hold, as described in 7.2. This determines the necessity of blocking methods as a way to optimize the calculation. For the JaroWinkler algorithm, it seems they have the fasted expected time to compute out of all. Therefore, JaroWinkler is a clear

winner in both abilities to identify matches and the shortest time required to perform these comparisons.

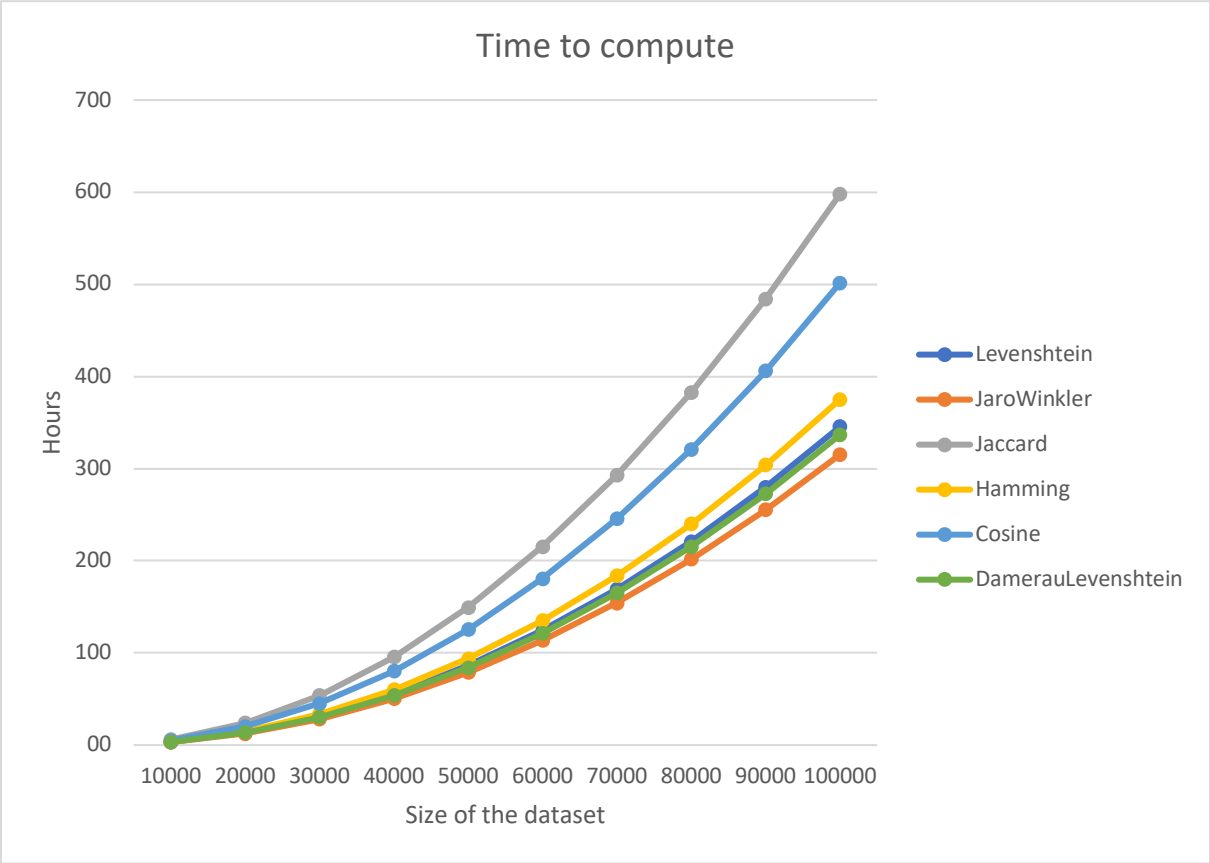

Figure 12 Estimated time to compute per algorithm (excluding SmithWaterman)

4.3.4.3 Discussion

An initial observation of the boxplot shows that the interquartile ranges are larger for all the addresses than those of the names. This indicates that the similarity of the names, for the matches, is overall much higher than the addresses. Addresses usually have longer strings than the name with abbreviations and additions. Abbreviations, such as “ave” and “avenue”, and additions, such as “21 street” and “21st street”, will reduce the similarity whilst still referring to the same thing. Therefore, we can derive from this data set that the similarity threshold of the address can be softer than that of the name.

JaroWinkler is the best recall performer from all the algorithms. We consider this the winner even though two other algorithms scored higher with the MCC score. Due to the decision to exclude all results lower than 0,1, it resulted in an impact on the number of true negatives. This does not impact recall, F-score or F2-score, but it does shift the result in favour of the Levensthein-based algorithms. If we included these filtered results, JaroWinkler would have been the best performer across all scorings.

All algorithms maintained perfect precision. Whilst we approached this from a statistical threshold point of view, it is recommended to further optimize the threshold and resulting F-score or the other scores. This could be done by feeling or systematically going through all thresholds to find the ideal balance of precision and recall. However, there can be

considerations to keep a precision high to minimize the number of false positives in case manual review is unfavourable or there is no risk appetite for having false positives.

From a time-to-compute perspective, the time numbers are projected on larger data sizes based on the performance of the RIDDLE data set. It shows that JaroWinkler is the fastest, but there is a clear threshold where using the algorithm without optimizers becomes impractical, i.e. more than 24 hours. This is purely from a time perspective, and there could be more limitations as the sizes increase and related requirements on hardware resources. Therefore, the size of the data set can be smaller before running into limitations to compute.

4.4 Blocking selection

We have identified three types of research that specifically mentioned and utilized a blocking technique within their integration activities. Due to the limited findings, we will collect proposals from the practitioners to gather experience- and pragmatic-based blocking methods to optimize performance.

4.4.1 Blocking techniques from the literature review

Sorted neighborhood indexing was an observed blocking technique for integration of US service member data [61]. This technique sorts all the different data values based on pre-determined criteria. The sorting allows for a search based on a window value set where the window is the number of neighbouring indices the subset will contain for comparison, i.e. the neighbourhood. This can be particularly powerful if you account for the data quality and characteristics of the data. For example, sorting strings in alphabetical order can provide strong similarity neighbourhoods if you expect typos. A set of records belonging to a single entity could be represented as this: [Philip, Philips, Phillips, Phlips]. Irrespective of the index, the neighbouring indices can be strong cases for matches, whilst the further indices (not shown in this example set) are less likely to be similar. Soundex was utilized on top of the strings to extract how the string is pronounced [61]. This allows for any (slight) string variation to be accounted for and group similar-sounding words. The previous example would become a set of the following Soundex keys: [P410, P412, P412, P412]. This gives a stronger initial indication that the last three strings are matches based on the Soundex alone.

A simpler method was also observed based on key attributes. The selection of a first name, last name, and an identifier such as a military service number or postal code was used for blocking [57], [61]. This relies on having high data quality and/or the assumption that these are strong features for identifying distinct matches. For example, typos or different standards in storing addresses can result in the exclusion of potential records from comparison or too familiar attribute selection can result in too broad inclusion of potential records. With additional preparation and insights into the characteristics of the data, there is a risk of having a narrower or too broad of a blocking result. However, this can be a deliberate reason for the organization to consider based on their needs.

The last blocking method that we observed was that use of a synonym dictionary [59]. In the context of integrating large bodies of text, certain (combinations of) strings were identified and checked whether or not they existed in a synonym dictionary. This resulted in an overview of known synonyms to the initial string. The research used these additional strings to check for occurrences in other texts. The research did not specifically mention this method to improve computing time; instead, they used it to widen the search. This depends on the implementation if it is a pure search through synonyms or having the synonyms trigger additional searches on top of your initial comparisons.

4.4.2 Principles from the practitioners

Due to the limited results on blocking within the literature review, we shared our findings with the practitioners. We identified priorities and requirements for using any technique that improves the time to compute. The input was gathered based on discussions with the BPEs

and BIEs from the organization regarding the use of any blocking method or optimizers to decrease the compute size. The results have been generalized into key principles to consider when considering and utilizing any optimization approach.

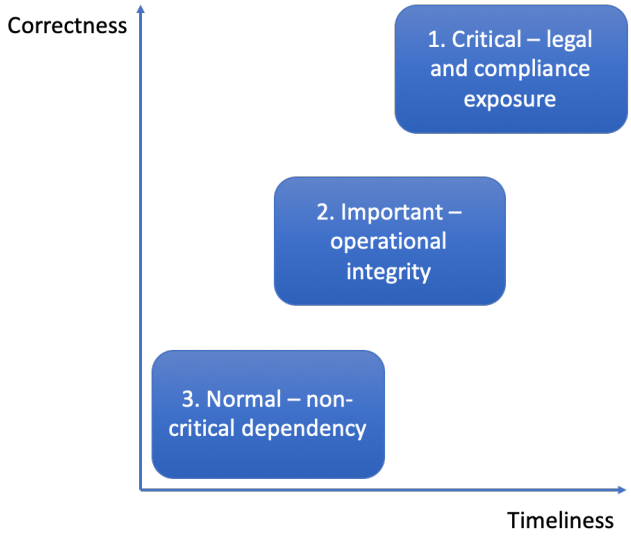

Figure 13 Levels of importance for optimizer consideration

Utilising a blocking method is not a trivial consideration in the organisational setting. From the literature review, the studies aimed to achieve integration through various techniques and concluded feasibility through qualitative statements or performance through quantitative evaluations. However, this was always performed on a homogeneous data set without consideration of importance. For the organisation, it is critical to distinguish between the types of data that they have: 1) critical, 2) important) and 3) normal. The exact criteria are dependent on the organisation, but having these three categories are required for the consideration of using a blocking technique.

The first category falls under a zero-tolerance policy where all activities to this data set occur in a controlled and versioned manner. Due to the highest level of legal and compliance scrutiny, this subset of data must achieve 100% correctness. This means that the integration algorithm must perform perfectly on precision and recall. Therefore, there is no room for any blocking method due to the potential risk of negative influence on those metrics. Consequently, there is an inverse relationship between the importance and the volume of data. In the case of the organisation, the volume of critical data does not pose any computational limitations.

The second category falls under strong importance due to the (immediate) impact it can have on operational activities. In the Purchase to Pay process, this negatively impacts the purchasing and manufacturing activities with financial consequences. The volume of this data set can be considerable, and coupled with the importance to business operations, there is a case to be made for utilising blocking methods. For the organisation, this category would require a minimum of weekly refresh cadence or, ideally, (near) real-time with high requirements for correct integration. The margin of error, dependent on the organisation, determines the time-to-compute gains it can achieve whilst maintaining high performance with a particular blocking method.

The third category falls under normal importance due to the need for business criticality due to their service offering or type of supplied goods. Generally, these cause inconveniences with no disruption to the core business. However, these are not of low importance due to (contractual) obligations; there is still the risk of financial consequences, making this category relevant to have integrated correctly.

Conclusively, the organisation prioritises the importance of the data and having it correct rather than purely computational performance-oriented. The discussion is about more than which blocking method should be utilised and how we can ensure no and minimal errors are made with the critical and important data, respectively. Using a blocking method can increase risk if not assessed carefully. For this organisation, the computer can run for up to a week, giving a decent margin before any need for blocking.

4.5 Summary of the results for sub-RQ1 and 2

This chapter provided the results to sub-research question 1 and 2. Figure 14 shows the key inputs for the design based on this chapter. For the first question, we found several integration algorithms that were researched but none of them were applied to Supplier Master Data. We found that Supplier Master Data shares characteristics with the generic Master Data domain in its purpose of being a key reference point across activities and processes. Additionally, it shares attribute characteristics with the Customer Domain such as name and address fields.

Researched integration algorithms from the overlapping domains were compared to identify which would perform the strongest on Supplier Master Data. The JaroWinkler algorithm showed the highest performance in accuracy and recall. This algorithm will be proposed in the artefact design and further validated.

Blocking techniques ideally provide a reduction in the number record comparisons which leads to faster results whilst maintaining performance. The techniques from the literature review were not utilized because the stakeholders opted for a risk-based blocking approach. Suppliers were categorized into critical, important, and normal. This led to the improvement of compute and also follow-up reviews to maximize performance based on the criticality of the supplier.

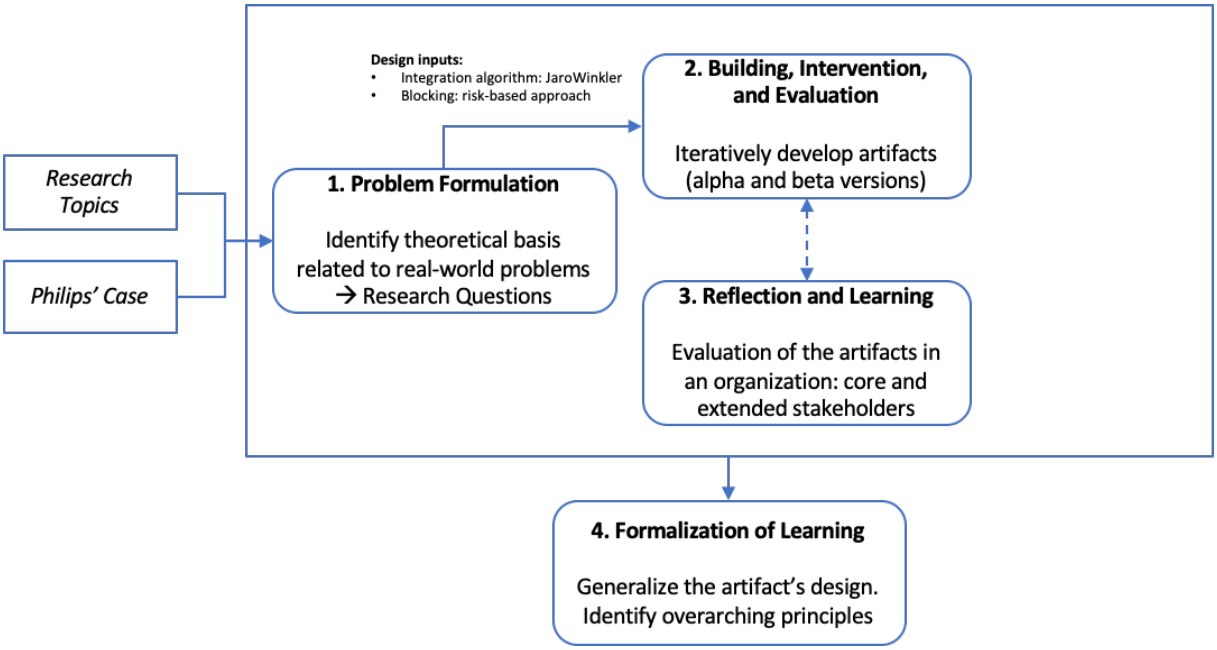

Figure 14 Initial design inputs from theory

5 Validation [REDACTED]

5.1 Proposed Artefact design [REDACTED]

The resulting artefact design that is subject to validation can be found in Figure 15. The overall setup includes three main components: 1) Data Sources, 2) Data Integration, and 3) Visualization & Consume. These are the minimal components necessary for the organization to validate integration artefacts. The first component is necessary for data integration as it provides the understanding of the data to be integrated at the attribute-, object- and architecture levels. The second component is the area of focus for this research and incorporates the results performed in previous chapters. The third component was not covered in the artefact design, as this serves mainly as a visualization layer for (non-technical) stakeholders to consume the results in the exact manner required.

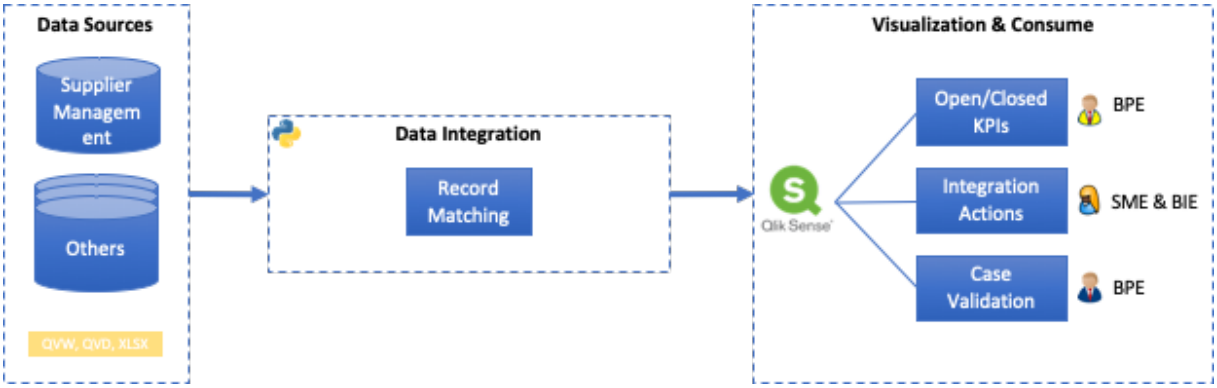

Figure 15 Integration solution design for validation

5.1.1 Data sources [REDACTED]

5.1.2 Data Integration [REDACTED]

5.1.3 Visualization & Consume [REDACTED]

5.2 Overview of validation strategy [REDACTED]

5.3 Results [REDACTED]

5.3.1 JaroWinkler algorithm [REDACTED]

5.3.2 Mean threshold classifier [REDACTED]

5.3.3 Blocking optimizer [REDACTED]

5.3.4 Case validation [REDACTED]

5.3.5 Other feedback [REDACTED]

5.3.5.1 Ranking rules and action identification [REDACTED]

5.3.5.2 Filtering [REDACTED]

5.3.5.3 Integration Actions database [REDACTED]

5.3.5.4 Supplier tracking [REDACTED]

5.4 Summary of the validation with the organization [REDACTED]

6 Updated Artefact Design [REDACTED]

7 Discussion and limitations [REDACTED]

7.1 Classifier optimization (precision or recall) [REDACTED]

7.2 Strong dependency on data quality and standardizations [REDACTED]

7.3 Create actionable and measurable insights [REDACTED]

7.4 Use of complete custom implementation versus vendor tool selection [REDACTED]

7.5 Limitations [REDACTED]

8 Conclusion [PARTIALLY REDACTED]

This thesis set out to solve a real-world problem experienced by Philips. The problem was defined as **supplier master data inconsistencies between multiple systems, which negatively impacted Purchase to Pay operations.** Action Design Research was used to find a solution. This methodology guided the design of an artefact that was based on theoretical knowledge with a strong focus on including inputs from practice. This artefact was required to integrate Supplier Master Data across various systems in Philips' IT-landscape. However, there was a research gap on Supplier Master Data integration and an important organizational requirement to minimise manual review. This research set out to identify the best approach to integrate Supplier Master Data and design an artefact that is cost-efficient for the organization.

Main RQ: How can Supplier Master Data be integrated for an organization? *The main research question is to address the problem statement experienced by Philips. This should include a holistic solution that addresses the problem through data integration and activities to sustain the solution, specifically in the Supplier Master Data domain.*

The main research is answered with the final artefact design described in Chapter 10. At a high level, it includes three components to achieve data integration: 1) data sources, 2) data integration, and 3) visualization & consumption. Each component has a critical dependency on each other as each component individually means nothing. The data integration component is only helpful in defining and preparing the data and having the correct supporting visualizations to consume it and validate its results. If done right, the crux remains within the data integration component, which we have further elaborated on in Figure 21. This includes all the necessary activities/capabilities to go from a set of supplier records to identifying the survivor and the loser(s). Merging the losers with the winner in all related data objects and systems achieves end-to-end integration.

Sub-RQ1: What integration algorithms can be applied to Supplier Master Data? *To identify integration algorithms applied in other data domains from literature and evaluate their applicability to Supplier Master Data.*

Initially, we have identified what characterizes the Supplier Master Data domain. This led to the description of Supplier Master Data based on similarity in attributes with existing Customer Master Domain data and overlap in the purpose of use from the general Master Data Domain. We have applied a comparison of integration algorithms that were researched in these domains. Except for machine learning approaches, we have measured the performance of edit-distance and token-based algorithms on precision, recall and variations of F-score. The JaroWinkler algorithm performed best across all metrics except for the MCC score, where Levenshtein was slightly higher. JaroWinkler was used in the artefact design and proved its validity based on quantitative and qualitative evaluation.

Sub-RQ2: How can an organization reduce manual review of integration results? *To identify activities in the solution that will reduce time-intensive manual review. This will provide pragmatic benefits as, for organizations, time is money.*

Two factors mainly contributed to the reduction of manual review: 1) increase of integration matching performance by the algorithm, i.e. higher precision and recall, or 2) risk-based prioritization. The first factor depends on the type of integration algorithm most suitable to the SMD, which was answered in Sub-RQ1. Additional performance can be gained by tuning the threshold and accounting for data quality issues. However, this never resulted in perfect and complete record matches. A precision of 0,8 and recall of 0,9 were achieved.

[REDACTED]

8.1 Contributions [PARTIALLY REDACTED]

We can summarize our contributions to the field of research and to the organization:

- **Research contributions:**
  - We provided a comparison of previously researched integration algorithms that can be grouped based on the type of data domain. This comparison was performed with a statistical threshold approach and on a data set that is publicly available. This allows for further comparisons of different integration algorithms and fair comparisons.
  - We have introduced a risk-based prioritization approach to minimize manual review based on the importance of suppliers. This is supplementary to blocking techniques, classification and data quality tuning that can be applied during record matching, i.e. improving performance.
  - [REDACTED]
- **Organizational contribution:**
  - [REDACTED]

8.2 Future work

We propose different areas that can be further explored and formally researched to:

- Ranking rules framework. The ranking rules component was critical as it applies a rule-based prioritization onto a group of matching records to which the record will remain. The potential here lies in guiding how to implement such rules and the possibility of providing a framework with predetermined rules that can be used for this purpose. Alternatively, this can be on a design principle level for research purposes or at the system/vendor level to maximally support organizational benefit.
- Use of machine learning as a classifier. Inclusion of machine learning elements to train the classifier within this design setup. As the critical category are all manually reviewed and the vital category sampled, these can provide a more extensive training and test set for training. This would make it more feasible to apply machine learning, given that the data also has enough attributes to derive features.
- Supplier tracking. The ability to track new and changed suppliers across systems can be further explored at a low cost. Specifically, which components are required and supporting processes to guide the artefact design? Additionally, there is potential to research how integration actions can be validated without implementing a large number of data checks and/or the inclusion of manual review. A dashboard helps

speed up the process, and building checks automates this process but can be time-consuming and lacks flexibility.

9 Bibliography

[1] S. Tuck, "Is MDM the route to the Holy Grail?," *Journal of Database Marketing & Customer Strategy Management*, vol. 15, no. 4, pp. 218–220, 2008.

[2] R. Silvola, O. Jaaskelainen, H. Kropsu-Vehkapera, and H. Haapasalo, "Managing one master data - Challenges and preconditions," *Industrial Management and Data Systems*, vol. 111, no. 1, pp. 146–162, 2011.

[3] A. Haug and J. S. Arlbjørn, "Barriers to master data quality," *Journal of Enterprise Information Management*, vol. 24, no. 3, pp. 288–303, 2011.

[4] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[5] A. Cleven and F. Wortmann, "Uncovering four strategies to approach master data management," *Proceedings of the Annual Hawaii International Conference on System Sciences*, no. October 2015, 2010.

[6] M. A. Jaro, "Probabilistic linkage of large public health data files," *Statistics in medicine*, vol. 14, pp. 491–498, 1995.

[7] P. Christen, *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. 2012.

[8] W. E. Winkler, "Overview of Record Linkage and Current Research Directions," *Bureau of the Census*, pp. 603–623, 2006.

[9] W. W. Cohen and S. E. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks," *IIWeb*, vol. 2003, pp. 73–78, 2003.

[10] R. C. Steorts, S. L. Ventura, M. Sadinle, and S. E. Fienberg, "A Comparison of Blocking Methods for Record Linkage," *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference*, pp. 1–22, 2014.

[11] H. Köpcke and E. Rahm, "Frameworks for entity matching: A comparison," *Data and Knowledge Engineering*, vol. 69, no. 2, pp. 197–210, 2010.

[12] G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, "Blocking and Filtering Techniques for Entity Resolution," *ACM Computing Surveys (CSUR)*, vol. 53, no. 2, pp. 1–42, 2020.

[13] P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 9, pp. 1537–1555, 2011.

[14] P. Christen, “A Comparison of Personal Name Matching : Techniques and Practical Issues,” *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW’06)*, pp. 290–294, 2006.

[15] L. K. Branting, “A Comparative Evaluation of Name-Matching Algorithms,” *Proceedings of the 9th international conference on Artificial intelligence and law*, pp. 224–232, 2003.

[16] P. Domingos and S. Parag, “Entity Resolution with Markov Logic,” *Sixth International Conference on Data Mining (ICDM’06)*, pp. 572–582, 2006.

[17] P. Christen and K. Goiser, “Quality and Complexity Measures for Data Linkage and Deduplication,” *Quality measures in data mining*, pp. 127–151, 2007.

[18] H. Köpcke, A. Thor, and E. Rahm, “Evaluation of entity resolution approaches on real-world match problems,” *Proceedings of the VLDB Endowment*, vol. 3, no. 1, pp. 484–493, 2010.

[19] D. G. Brizan and A. U. Tansel, “A Survey of Entity Resolution and Record Linkage Methodologies,” *Communications of the IIMA*, vol. 6, no. 3, pp. 41–50, 2006.

[20] Y. Zhou and J. R. Talburt, “Entity Identity Information Management (EIIM),” *ICIQ*, 2011.

[21] D. Mahata and J. Talburt, “A framework for collecting and managing entity identity information from social media,” *Proceedings of the 19th International Conference on Information Quality, ICIQ 2014*, pp. 216–233, 2014.

[22] S. Ben Hassine-Guetari and B. Laboisse, “Managing multisource databases: Between theory and practice,” *ICIQ*, 2011.

[23] C. Vercellis, “Business Intelligence: Data Mining and Optimization for Decision Making,” *Business Intelligence: Data Mining and Optimization for Decision Making*, pp. 1–417, 2009.

[24] L. Cai and Y. Zhu, “The challenges of data quality and data quality assessment in the big data era,” *Data Science Journal*, vol. 14, pp. 1–10, 2015.

[25] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature Reviews in Software Engineering,” 2007.

[26] L. Shamseer *et al.*, “Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: Elaboration and explanation,” *BMJ (Online)*, vol. 349, no. February 2011, pp. 1–25, 2015.

[27] F. Haneem, N. Kama, R. Ali, and A. Selamat, “Applying data analytics approach in

systematic literature review: Master data management case study," *Frontiers in Artificial Intelligence and Applications*, vol. 297, no. September, pp. 705–715, 2017.

[28] L. H. Hansen, R. Van Son, A. Wieser, and E. Kjems, "ADDRESSING the ELEPHANT in the UNDERGROUND: AN ARGUMENT for the INTEGRATION of HETEROGENEOUS DATA SOURCES for RECONCILIATION of SUBSURFACE UTILITY DATA," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 46, no. 4/W4-2021, pp. 43–48, 2021.

[29] K. P. Kusuma Dewi, T. Fabrianti Kusumasari, and R. Andreswari, "Analysis and design of architecture master data management (MDM) tools for open source platform at PT XYZ," *Proceedings - 2019 5th International Conference on Science and Technology, ICST 2019*, pp. 244–249, 2019.

[30] J. Sreemathy, K. Naveen Durai, E. Lakshmi Priya, R. Deebika, K. Suganthi, and P. T. Aisshwarya, "Data Integration and ETL: A Theoretical Perspective," *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, pp. 1655–1660, 2021.

[31] M. A. Naeem, G. Dobbie, G. Weber, P. Bag, P. Street, and N. Zealand, "An Event-Based Near Real-Time Data Integration Architecture," *Enterprise Distributed Object Computing Conference Workshops*, vol. 12, pp. 10–13, 2008.

[32] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, "Combining actual trends in software systems for business management," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 313–324, 2003.

[33] K. Murthy *et al.*, "Exploiting evidence from unstructured data to enhance master data management," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1862–1873, 2012.

[34] M. S. Sezgin, A. T. Bayrak, and O. T. Ytldtz, "A Hybrid Approach to Dynamic Enterprise Data Platform," *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 3486–3492, 2019.

[35] D. Jaksic, V. Jovanovic, and P. Poscic, "Integrating evolving MDM and EDW systems by data vault based system catalog," *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings*, pp. 1401–1406, 2017.

[36] D. Kim *et al.*, "System architecture and information model for integrated access to distributed biomedical information," *Medical Imaging 2009: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, vol. 7264, no. March 2009, p. 72640G, 2009.

[37] P. Title, “Paper Title : - Master Data Management-CDI,” *Conrad, J. G., Dozier, C., Molina-Salgado, H., Thomas, M., & Veeramachaneni, S. (2011). Public record aggregation using semi-supervised entity resolution. Proceedings of the International Conference on Artificial Intelligence and Law, 239–248. https://doi*, no. November, 2008.

[38] L. Menet, M. Lamolle, and A. Zerdazi, “Managing master data with XML schema and UML,” *Proceedings - International Workshop on Advanced Information Systems for Enterprises, IWAISE 2008*, pp. 53–59, 2008.

[39] M. A. Naeem, G. Dobbie, I. S. Bajwa, and G. Weber, “Resource optimization for processing of stream data in data warehouse environment,” *ACM International Conference Proceeding Series*, pp. 62–68, 2012.

[40] M. A. Naeem, G. Dobbie, and G. Weber, “Efficient processing of streaming updates with archived master data in near-real-time data warehousing,” *Knowledge and Information Systems*, vol. 40, no. 3, pp. 615–637, 2014.

[41] C. Chirathamjaree, “A data model for heterogeneous data sources,” *IEEE International Conference on e-Business Engineering, ICEBE’08 - Workshops: AiR’08, EM2I’08, SOAIC’08, SOKM’08, BIMA’08, DKEEE’08*, pp. 121–127, 2008.

[42] D. Hutchison and J. C. Mitchell, *Business intelligence for the real-time enterprises: first international workshop, BIRTE 2006*, vol. 4365. 2006.

[43] B. Piprani, “A model for semantic equivalence discovery for harmonizing master data,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5872 LNCS, pp. 649–658, 2009.

[44] S. Nadal, A. Abelló, O. Romero, S. Vansummeren, and P. Vassiliadis, “MDM: Governing evolution in big data ecosystems,” *Advances in Database Technology - EDBT*, vol. 2018-March, pp. 682–685, 2018.

[45] M. Dugas, A. Meidt, P. Neuhaus, M. Storck, and J. Varghese, “ODMedit: Uniform semantic annotation for data integration in medicine based on a public metadata repository,” *BMC Medical Research Methodology*, vol. 16, no. 1, pp. 1–9, 2016.

[46] M. Dugas *et al.*, “Compatible data models at design stage of medical information systems: Leveraging related data elements from the MDM portal,” *Studies in Health Technology and Informatics*, vol. 264, pp. 113–117, 2019.

[47] M. K. Sein, O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren, “Action design research,” *MIS Quarterly: Management Information Systems*, vol. 35, no. 1, pp. 37–56, 2011.

[48] R. Wieringa, *Design Science Methodology for Information Systems and Software Engineering*. 2014.

[49] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.

[50] S. Al-Janabi and R. Janicki, "A density-based data cleaning approach for deduplication with data consistency and accuracy," *Proceedings of 2016 SAI Computing Conference, SAI 2016*, pp. 492–501, 2016.

[51] M. Neuhaus and H. Bunke, "Edit distance-based kernel functions for structural pattern classification," *Pattern Recognition*, vol. 39, no. 10, pp. 1852–1863, 2006.

[52] A. Bookstein, V. A. Kulyukin, and T. Raita, "Generalized hamming distance," *Information Retrieval*, vol. 5, no. 4, pp. 353–375, 2002.

[53] R. Haldar and D. Mukhopadhyay, "Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach," *arXiv preprint arXiv:1101.1232*, no. Ld, 2011.

[54] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information (Switzerland)*, vol. 11, no. 9, pp. 1–17, 2020.

[55] J. Wang, G. Li, and J. Fe, "Fast-join: An efficient method for fuzzy token matching based string similarity join," *Proceedings - International Conference on Data Engineering*, pp. 458–469, 2011.

[56] S. Karpischek, F. Michahelles, and E. Fleisch, "Detecting incorrect product names in online sources for product master data," *Electronic Markets*, vol. 24, no. 2, pp. 151–160, 2014.

[57] J. G. Conrad, C. Dozier, H. Molina-Salgado, M. Thomas, and S. Veeramachaneni, "Public record aggregation using semi-supervised entity resolution," *Proceedings of the International Conference on Artificial Intelligence and Law*, pp. 239–248, 2011.

[58] K. Chhaya, A. Khanzode, and R. D. Sarode, "Advantages and disadvantages of artificial intelligence and machine learning: A literature review," *International Journal of Library & Information Science (IJLIS)* , vol. 9, no. 1, pp. 30–36, 2020.

[59] A. Dutta, T. Deb, and S. Pathak, "Automated Data Harmonization (ADH) using Artificial Intelligence (AI)," *Opsearch*, vol. 58, no. 2, pp. 257–275, 2021.

[60] D. Deng *et al.*, "Unsupervised string transformation learning for entity consolidation," *Proceedings - International Conference on Data Engineering*, vol. 2019-April, pp. 196–

207, 2019.

[61] J. D. Warnke-Sommer and F. E. Damann, “An improved machine learning application for the integration of record systems for missing US service members,” *International Journal of Data Science and Analytics*, vol. 11, no. 1, pp. 57–68, 2021.

[62] E. Rahm and H. H. Do, “Data Engineering - Special Issue on Data Cleaning,” *Data Engineering*, vol. 23, no. 4, pp. 3–13, 2000.

10 Appendix

Title	Q1	Q2	Q3
A data model for heterogeneous data sources	1	1	1
A density-based data cleaning approach for deduplication with data consistency and accuracy	1	1	1
A holistic approach for the architecture and design of an ontology-based data integration capability in product master data management	1	0,5	0
A Hybrid Approach to Dynamic Enterprise Data Platform	0,5	1	1
A model for semantic equivalence discovery for harmonizing master data	0,5	1	1
ADDRESSING the ELEPHANT in the UNDERGROUND: AN ARGUMENT for the INTEGRATION of HETEROGENEOUS DATA SOURCES for RECONCILIATION of SUBSURFACE UTILITY DATA	0,5	1	1
An event-based near real-time data integration architecture	1	1	0
An improved machine learning application for the integration of record systems for missing US service members	1	0	1
An Industrial Dynamic Skyline Based Similarity Joins for Multidimensional Big Data Applications	0,5	1	1
Analysis and Design of Data Synchronization Algorithm for Master Data Management Tools Based on Open Source Platform at PT. XYZ	0,5	0	1
Automated Data Harmonization (ADH) using Artificial Intelligence (AI)	0,5	1	1

BayesWipe: A multimodal system for data cleaning and consistent query answering on structured bigdata	1	1	1
Callisto: Mergers without pain	1	1	1
Combining actual trends in software systems for business management	0	1	0
Compatibility between metadata standards: Import pipeline of CDISC ODM to the samply.MDR	0	0	0
Compatible data models at design stage of medical information systems: Leveraging related data elements from the MDM portal	1	1	1
Corroborating quality of data through density information	1	1	1
Data Integration and ETL: A Theoretical Perspective	0	1	0
Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers	1	0	1
Detecting incorrect product names in online sources for product master data	1	1	1
Efficient processing of streaming updates with archived master data in near-real-time data warehousing	1	1	1
Exploiting evidence from unstructured data to enhance master data management	1	1	1
Integrating evolving MDM and EDW systems by data vault based system catalog	1	1	1
Managing master data with XML schema and UML	0	1	0
Master data management-CDI	1	0	1

MDM: Governing evolution in big data ecosystems	0	0	1
ODMedit: Uniform semantic annotation for data integration in medicine based on a public metadata repository	1	1	1
Public record aggregation using semi-supervised entity resolution	1	1	1
Resource optimization for processing of stream data in data warehouse environment	1	1	1
System architecture and information model for integrated access to distributed biomedical information	0	0	0
TPCDI: The first industry benchmark for data integration	1	1	1
Unsupervised string transformation learning for entity consolidation	1	1	1

Table 8 Quality assessment results

Title	Indexing / Blocking	Edit Distance Comparison	Token Based Comparison	Supervised ML - Classification"	Unsupervised ML - Clustering
A density-based data cleaning approach for deduplication with data consistency and accuracy					Density-based weight model
A Hybrid Approach to Dynamic Enterprise Data Platform					
An improved machine learning application for the integration of record systems for missing US service members	Sorted neighborhood indexing based on soundex encoding Sorted neighborhood indexing basedon key identifier (service number)	Damerau Levenshtein	Cosine Similarty	Naïve Bayesian Logistic Regression Support Vector Machine	
An Industrial Dynamic Skyline Based Similarity Joins for Multidimensional Big Data Applications					

Title	Indexing / Blocking	Edit Distance Comparison	Token Based Comparison	Supervised ML - Classification"	Unsupervised ML - Clustering
Automated Data Harmonization (ADH) using Artificial Intelligence (AI)	Synonym search in a dictionary	Jaccard	Cosine Similarity		K-Means Binary Classifier Bayesian Support Vector Machine Random Forest Ada Boost
BayesWipe: A multimodal system for data cleaning and consistent query answering on structured bigdata					Bayesian Network
Corroborating quality of data through density information					Density Based Model
Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers					

Title	Indexing / Blocking	Edit Distance Comparison	Token Based Comparison	Supervised ML - Classification"	Unsupervised ML - Clustering
Detecting incorrect product names in online sources for product master data		Equality Levenshtein Jaro-Winkler Word Coefficient	Q-Grams*	Naïve Bayes Classification Tree Support Vector Machine Logistc Regression	
Public record aggregation using semi-supervised entity resolution	Blocking key used on first name, last name and first digits of the postal code	Hamming Levenshtein Jaro-Winkler	Cosine Similarity Smith-Waterman TD-IDF*	Support Vector Machine	
Unsupervised string transformation learning for entity consolidation					Custom

Table 9 Overview of research and integration algorithm

Title	Data Domain	Data Sub-Domain
A data model for heterogeneous data sources	Master Data	-
A density-based data cleaning approach for deduplication with data consistency and accuracy	People Data	Customer
A Hybrid Approach to Dynamic Enterprise Data Platform	People Data	Customer
A model for semantic equivalence discovery for harmonizing master data	Master Data	Aviation
ADDRESSING the ELEPHANT in the UNDERGROUND: AN ARGUMENT for the INTEGRATION of HETEROGENEOUS DATA SOURCES for RECONCILIATION of SUBSURFACE UTILITY DATA	Master Data	Geolocation
An event-based near real-time data integration architecture	Master Data	-
An improved machine learning application for the integration of record systems for missing US service members	People Data	Employee
An Industrial Dynamic Skyline Based Similarity Joins for Multidimensional Big Data Applications	Big Data	-
Analysis and Design of Data Synchronization Algorithm for Master Data Management Tools Based on Open Source Platform at PT. XYZ	Master Data	-

Title	Data Domain	Data Sub-Domain
Automated Data Harmonization (ADH) using Artificial Intelligence (AI)	Master Data	Product
BayesWipe: A multimodal system for data cleaning and consistent query answering on structured bigdata	Big Data	-
Callisto: Mergers without pain	Master Data	Product and Customer
Combining actual trends in software systems for business management	Master Data	-
Compatible data models at design stage of medical information systems: Leveraging related data elements from the MDM portal	People Data	Medical
Corroborating quality of data through density information	People Data	Customer
Data Integration and ETL: A Theoretical Perspective	-	-
Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers	People Data	Medical
Detecting incorrect product names in online sources for product master data	Master Data	Product

Title	Data Domain	Data Sub-Domain
Efficient processing of streaming updates with archived master data in near-real-time data warehousing	Big Data	-
Exploiting evidence from unstructured data to enhance master data management	Master Data	Unstructured
Integrating evolving MDM and EDW systems by data vault based system catalog	People Data	Employee
Managing master data with XML schema and UML	Master Data	-
Master data management-CDI	People Data	Customer
MDM: Governing evolution in big data ecosystems	Big Data	-
ODMedit: Uniform semantic annotation for data integration in medicine based on a public metadata repository	People Data	Medical
Public record aggregation using semi-supervised entity resolution	People Data	General individuals and public records
Resource optimization for processing of stream data in data warehouse environment	Big Data	-

Title	Data Domain	Data Sub-Domain
System architecture and information model for integrated access to distributed biomedical information	People Data	Medical
TPCDI: The first industry benchmark for data integration	People Data	Customer
Unsupervised string transformation learning for entity consolidation	People Data	General individuals and public records

Table 10 Data Domains

Title	Case Implementation	Artefact Development	Precision and recall	Scalability and runtime	Comparative Study	Filtering Ratio
A data model for heterogeneous data sources	1					
A density-based data cleaning approach for deduplication with data consistency and accuracy			1	1	1	
A Hybrid Approach to Dynamic Enterprise Data Platform	1					
A model for semantic equivalence discovery for harmonizing master data	1					
ADDRESSING the ELEPHANT in the UNDERGROUND: AN ARGUMENT for the INTEGRATION of HETEROGENEOUS DATA SOURCES for RECONCILIATION of SUBSURFACE UTILITY DATA	1					
An event-based near real-time data integration architecture						

Title	Case Implementation	Artefact Development	Precision and recall	Scalability and runtime	Comparative Study	Filtering Ratio
An improved machine learning application for the integration of record systems for missing US service members			1	1		
An Industrial Dynamic Skyline Based Similarity Joins for Multidimensional Big Data Applications				1	1	1
Analysis and Design of Data Synchronization Algorithm for Master Data Management Tools Based on Open Source Platform at PT. XYZ	1					
Automated Data Harmonization (ADH) using Artificial Intelligence (AI)			1	1		
BayesWipe: A multimodal system for data cleaning and consistent query answering on structured bigdata			1	1		
Callisto: Mergers without pain	1	1				

Title	Case Implementation	Artefact Development	Precision and recall	Scalability and runtime	Comparative Study	Filtering Ratio
Combining actual trends in software systems for business management						
Compatible data models at design stage of medical information systems: Leveraging related data elements from the MDM portal		1				
Corroborating quality of data through density information			1	1	1	
Data Integration and ETL: A Theoretical Perspective						
Data Integration Protocol In Ten-steps (DIPIT): A new standard for medical researchers	1					
Detecting incorrect product names in online sources for product master data			1			
Efficient processing of streaming updates with archived master data in near-real-time data warehousing				1		

Title	Case Implementation	Artefact Development	Precision and recall	Scalability and runtime	Comparative Study	Filtering Ratio
Exploiting evidence from unstructured data to enhance master data management			1	1		
Integrating evolving MDM and EDW systems by data vault based system catalog	1					
Managing master data with XML schema and UML						
Master data management-CDI		1				
MDM: Governing evolution in big data ecosystems	1					
ODMedit: Uniform semantic annotation for data integration in medicine based on a public metadata repository	1	1				
Public record aggregation using semi-supervised entity resolution			1			

Title	Case Implementation	Artefact Development	Precision and recall	Scalability and runtime	Comparative Study	Filtering Ratio
Resource optimization for processing of stream data in data warehouse environment				1		
System architecture and information model for integrated access to distributed biomedical information	1	1				
TPCDI: The first industry benchmark for data integration			1	1		
Unsupervised string transformation learning for entity consolidation			1			

Table 11 Evaluation methods

Data Domain	Edit Distance Comparison	Token Based Comparison	Supervised ML - Classification"	Unsupervised ML - Clustering	Other
Big Data	-	-	-	Bayesian Network	Dynamic Skyline Query Join Operation
Master Data	Equality Levenshtein Jaccard Jaro-Winkler Word Coefficient	Q-Grams Cosine Similarity	Naïve Bayes Classification Tree Support Vector Machine Logistc Regression	K-Means Binary Classifier Bayesian Support Vector Machine Random Forest Ada Boost	-
People Data	Damerau Levenshtein Hamming Levenshtein Smith-Waterman Jaro-Winkler	Cosine Similarty	Naïve Bayesian Logistic Regression Support Vector Machine	Density Based Model	File Integration Strategy

Table 12 Overview of integration algorithms per domain